On the Automatic Recognition of Facial Non-verbal Communication Meaning in Informal, Spontaneous Conversation

Tim Sheerman-Chase

Submitted for the Degree of
Doctor of Philosophy
from the
University of Surrey

Centre for Vision, Speech and Signal Processing
Faculty of Engineering and Physical Sciences
University of Surrey
Guildford, Surrey GU2 7XH, U.K.

September 2012

Summary

Non-Verbal Communication (NVC) comprises all forms of inter-personal communication, apart from those that are based on words. NVC is essential to understand communicated meaning in common social situations, such as informal conversation. The expression and perception of NVC depends on many factors, including social and cultural context.

The development of methods to automatically recognise NVC enables new, intuitive computer interfaces for novel applications, particularly when combined with emotion or natural speech recognition. This thesis addresses two questions: *how can facial NVC signals be automatically recognised, given cultural differences in NVC perception?* and, *what do automatic recognition methods tell us about facial behaviour during informal conversations?*

A new data set was created based on recordings of people engaged in informal conversation. Minimal constraints were applied during the recording of the participants to ensure that the conversations were spontaneous. These conversations were annotated by volunteer observers, as well as paid workers via the Internet. This resulted in three sets of culturally specific annotations based on the geographical location of the annotator (Great Britain, India, Kenya). The cultures differed in the average label that the culture's annotators assigned to each video clip. Annotations were based on four NVC signals: agreement, thinking, questioning and understanding, all of which commonly occur in conversations. An automatic NVC recognition system was trained based on culturally specific annotation data and was able to make predictions that reflected cultural differences in annotation.

Various visual feature extraction methods and classifiers were compared to find an effective recognition approach. The problem was also considered from the perspective of regression of dimensional, continuous valued annotation labels, using Support Vector Regression (SVR), which enables the prediction of labels which have richer information content than discrete classes. The use of Sequential Backward Elimination (SBE) feature selection was shown to greatly increase recognition performance.

With a method for extracting the relevant facial features, it becomes possible to investigate human behaviour in informal conversation using computer tools. Firstly, the areas of the face used by the automatic recognition system can be identified and visualised. The involvement of gaze in thinking is confirmed, and a new association between gestures and NVC are identified, i.e. brow lowering (AU4) during questioning. These findings provide clues as to the way humans perceive NVC. Secondly, the existence of coupling in human expression is quantified and visualised. Patterns exist in both mutual head pose and in the mouth area, some of which may relate to mutual smiling. This coupling effect is used in an automatic NVC recognition system based on backchannel signals.

Key words: Non-verbal Communication, Annotation, Automatic Recognition, Cultural Factors, Subjective Labels, Crowdsourcing

Email: t.sheerman-chase@surrey.ac.uk

WWW: http://www.surrey.ac.uk/feps/

The living thing did I follow; I walked in the broadest and narrowest paths to learn its nature.

With a hundred-faced mirror did I catch its glance when its mouth was shut, so that its eye might speak unto me. And its eye spake unto me.

Friedrich Nietzsche

Destroy the mind, destroy the body, but you cannot destroy the heart

The Smashing Pumpkins

Acknowledgements

This work was a challenge to bring to fruition. I was supported and encouraged by many people and I thank them all. While it is impossible list everyone who assisted me, I will mention of a few special people.

To my immediate research team: Andy, Brian, Dumebi, Helen, James, Karel, Liam, Matt, Matt, Nico, Phil, Segun, Simon and Stephen, thanks for our many discussions on technical matters, current affairs and for good times. Also, I am very grateful for the friendship and encouragement of Ash. My research was initiated from the EPSRC funded LiLiR project and I want to extend my gratitude to Richard, Stephen, Barry, Gari, Yuxuan, Jake and Sarah for their useful input and providing direction to this work. I also wish to thank the university staff who assisted this work, particularly Bevis King, James Field, Liz James, Dawn Duke and Shane Dowle.

Thanks goes to my family and particularly to my parents for providing support that cannot be measured. Also, thanks to my old friends (you know who you are, I hope) and new friends Martin, Cemre, Stuart, Rui, Iris, Phil for providing a spot of sunshine. Many professionals helped me keep mind and body together; special thanks to Chris Powell, Philip Bull, Duncan Sellers, Nicole Muller, Dee Ainsworth and Brenda Dilks for your timely help.

Many thanks to Rich Bowden for giving me the opportunity to do research, for providing copious guidance for my own good (most of which I followed) and particularly for how to best communicate my research. His support, persistence and patience during my active work and during my absences was invaluable. Eng-Jon Ong was my scientific guru for the past 5 years and this work would have been impossible without his help.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

c	Non-Verbal Communication (NVC) category
d	Number of annotators
e	Number of temporal windows
f	Frame number
g	LBP histogram grid width
h	LBP histogram grid height
k	Temporal Window size
m	Clip ID number
o	Number of videos in the corpus
p	Number of frames in original video
q	Number of discretisation thresholds
r	Number of frames in a video clip
s	Number of feature components
t	Time offset
u	Clip digest feature value
w	Threshold for trusted worker subset

$\mathbf{d}$	Combined ratings for a single annotator
$\mathbf{e}$	Number of clips rated by each annotator
$\mathbf{f}$	Feature vector based on a single video frame
$\mathbf{g}$	Polynomial curve constants
$\mathbf{m}$	Mean value of feature components
$\mathbf{n}$	Tracker relevance weights
$\mathbf{o}$	Performance increase during feature selection
$\mathbf{p}$	Performance during feature selection
$\mathbf{r}$	Annotator rating of a single video clip
$\mathbf{s}$	Annotator index of clip rating
$\mathbf{t}$	Head translation
$\mathbf{v}$	Variance value of feature components
$\mathbf{z}$	Number of annotators for each clip

C	Cultural sets of annotators
E	Set of intense NVC examples
L	Length of video clip
V	Set of all video clips
N	Set of NVC categories
P	Quadratic polynomial function
W	Set of annotators

$\mathbf{A}$	Affine transform to frontal pose
$\mathbf{B}$	Feature matrix for a video clip
$\mathbf{C}$	Matrix of filtered consensus annotations
$\mathbf{D}$	Correlation of pairs of features
$\mathbf{E}$	Digest feature matrix for a clip
$\mathbf{F}$	Feature matrix for original video
$\mathbf{G}$	Feature matrix of concatenated original videos
$\mathbf{H}$	Tracker positions in homogeneous coordinates
$\mathbf{I}$	Filtered annotation data
$\mathbf{J}$	Polynomial curve feature matrix
$\mathbf{K}$	Cross cultural mean of annotation data
$\mathbf{L}$	Variance of features in a sliding window
$\mathbf{M}$	Normalised tracking
$\mathbf{N}$	Raw annotation data
$\mathbf{O}$	Discretisation thresholds
$\mathbf{P}$	Relevance weight of trackers
$\mathbf{Q}$	Discretised clip feature matrix
$\mathbf{R}$	Head rotation matrix
$\mathbf{S}$	Statistical feature matrix from multiple windows
$\mathbf{T}$	Tracking position data
$\mathbf{U}$	Consensus ratings of trusted annotators
$\mathbf{V}$	NVC ratings from a single annotator
$\mathbf{W}$	Feature matrix inside temporal window

$\boldsymbol{\alpha}$	Current set of feature components
$\boldsymbol{\beta}$	Feature component set for testing
ϵ	Number of frames in the corpus
η	Number of features to remove in iteration
κ	Number of trackers
μ	Statistical mode function
ν	Variable for correlation computation
ρ	Pearson's correlation function
τ	Camera projection function
ϕ	Variance function
$\boldsymbol{\chi}$	Tracker positions in homogeneous coordinates
ψ	Model fit error
ω	Number of feature components in set

AAM	Active Appearance Model
ASM	Active Shape Model
AU	Action Unit
AUC	Area Under Curve
CLM	Constrained Local Model
CORF	Conditional Ordinal Random Field
CRF	Conditional Random Field
DBN	Dynamic Bayesian Network
DCT	Discrete Cosine Transform
EM	Expectation Maximization
FACS	Facial Action Coding System
FLD	Fisher Linear Discriminant
GBR	Great Britain
HCI	Human–Computer Interaction
HCRF	Hidden Conditional Random Field
HMM	Hidden Markov Model
ICA	Independent Component Analysis
iHCRF	Infinite Hidden Conditional Random Field
IP	Internet Protocol
KLT	Kanade-Lucas-Tomasi
kNN	k-Nearest Neighbour
LBP	Local Binary Pattern
LDA	Linear Discriminant Analysis
LGBP	Local Gabor Binary Pattern
LLR	Logistical Linear Regression
LMA	Levenberg–Marquardt Algorithm
LP	Linear Predictor
LPQ	Local Phase Quantization
MIL	Multiple Instance Learning
MKL	Multiple Kernel Learning
ML	Machine Learning
MNN	Modified Nearest Neighbour
NMF	Non-negative Matrix Factorisation
NVC	Non-Verbal Communication
PCA	Principal Component Analysis
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
RVM	Relevance Vector Machine
SAL	Sensitive Artificial Listener
SBE	Sequential Backward Elimination
SFS	Sequential Forward Selection
SIFT	Scale-invariant Feature Transform
SVDD	Support Vector Data Description
SVM	Support Vector Machine
SVR	Support Vector Regression
TOP	Three Orthogonal Planes
UK	United Kingdom
US	United States

The study of Expression is difficult, owing to the movements being often extremely slight, and of a fleeting nature.

Charles Darwin [71]

1

Introduction

Non-Verbal Communication (NVC) comprises all forms of intentional, inter-personal communication apart from word based messages and is essential to understand communicated meaning in social situations [166] (see Figures 1.1 and 1.2 for illustrative examples)[1]. To enable new ways of intuitively interacting with computers, it would be beneficial for these devices to understand human communication, including NVC e.g. using innate communication skills to interact with computer characters. However, this is challenging because human behaviour expression and perception depends on many factors. This thesis describes a study based on the recording and culturally specific[2] annotation of videos of informal spontaneous conversations, and develops techniques

[1] This definition, together with the alternative perspective of NVC that is independent of intentionality is further discussed in Section 2. However, this is a good working definition.

[2] "culturally specific" refers to the annotation data being collected from distinct cultural groups of annotators. In this study, the expressors of NVC were not divided into distinct cultural groups.

Figure 1.1: Hand gestures and facial expressions are often used to convey NVC. The left photograph shows a woman waving goodbye and smiling. The right photograph depicts two armed security guards with their hands held in a blocking fashion while maintaining eye contact. Photographs used by permission (see Appendix C).

to successfully recognise meaningful communication based on visual non-verbal signals. The use of culturally specific, dimensional, continuous valued labels based on dimensional annotations allows techniques to more faithfully model the culturally specific NVC perception of the human observers. "Continuous value", in this case meaning "having an infinite number of possible values" rather than in the sense of "continuous recognition" meaning "providing a series of predictions based on a temporal series of observations". [3] This results in an approach that can recognise NVC signals which is a step towards a system that can be used via common and intuitive communication skills. The four NVC signals selected for study were *agree*, *understand*, *thinking* and *question*. This thesis addresses two questions: *how can facial NVC signals be automatically recognised, given cultural differences in NVC perception?* and, *what do automatic recognition methods tell us about facial behaviour during informal conversations?*

1.1 Motivation: Why Attempt Automatic Recognition of NVC?

Computers are becoming pervasively used in society because of their portability, connectivity and low cost. They are being integrated into many common home appliances,

[3] This is discussed further in Section 3.3.

including cars, mobile phones and televisions. NVC is a communication skill that almost all humans possess and may be used to operate computers if the appropriate interfaces are available. Furthermore, if human to human communication can be understood by a device, it may be possible for it to support or analyse the interaction [219]. Existing Human–Computer Interaction (HCI) interfaces do not necessarily extend to multiple embedded computers, and particularly to those without traditional physical input controls [237]. Potential applications of automatic NVC recognition and emotionally aware computing include:

- Computer generated characters and robots capable of a fully interactive conversation, including perception and expression of NVC and emotion [236]. Early prototypes of this concept include the SEMAINE Sensitive Artificial Listener [279] and Asimo the robotic museum guide [4]. The computer character could be used for entertainment, companionship or for providing information and assistance.

- Monitoring of social human behaviour and automatically providing services which are useful and appropriate for the social context. This could be in the form of advice, information or assistance [177]. Computer may be able to optimise work patterns and reduce untimely interruptions [143]. Emotionally aware computers may also provide an opportunity to practice social skills [141].

- Computer based learning can detect if a human is confused or bored and adapt the teaching style accordingly [82].

- Human operator monitoring in safety critical situations can detect tiredness, confusion or ignoring important information. This can be applied to transportation (cars [153], aircraft, etc.), industrial machinery or any safety critical device.

- Text entry based on audio-visual speech recognition. Because of the association between words, emotions and NVC [65, 156], a hybrid recognition approach could improve speech recognition performance.

- Automatic labelling of broadcast material, based on the behaviour of the people featured in the footage, or from the reaction of observers [235].

Applications of automatic NVC and behaviour recognition systems, in which the result is directly interpreted by humans include:

- Product evaluation. As a person interacts with a product, their reactions, emotions, mental state and intentions can be automatically recognised which provides useful information for evaluating a product's appeal [194].
- Security applications. Behaviour monitoring [249], covert lip reading [232] and lie detection can be used by police for crime detection. The work contained in this thesis originated from the EPSRC LILiR project, which developed video based lip reading technology.
- Human behaviour research. Higher level patterns in behaviour can be quickly analysed if low level NVC behaviour can be automatically recognised [230].

Automatic NVC recognition has a wide range of potential applications but these are often in roles that are quite different from existing computer systems. If human behaviour is to be automatically recognised, the issue of context and situational dependence of human behaviour must be considered.

1.2 From the Laboratory to the World (or Why Automatic NVC Recognition is Difficult)

The way NVC is expressed and perceived depends on the context in which it is used e.g. certain types of behaviours may be considered as inappropriate in specific social situations and not be expressed, as well as behaviour such as nodding, winking, kissing, nodding, etc. that are interpreted depending on the social relationship[4]. Many studies of automatic human behaviour recognition use posed data recorded in a laboratory environment [236, 121, 324]. Posed NVC data differs significantly from natural data. These differences include the types of communication actions used, expression intensity, the style of expression and in the timings of gestures. To maximise the range of

[4]This is discussed more fully in Section 2.1.

Figure 1.2: Eye contact has a significant role in NVC. The left photograph shows a mother and child during mutual eye contact. The proximity and body pose of the people implies a close intimacy. The right photograph shows a man winking, which is often intended as a sign of trust. Photographs used by permission (see Appendix C).

applications in which the automatic system can be applied, this thesis attempts to use spontaneous human behaviour where possible.

At the time this work was conducted, there were limited available public data sets that would support the research. Data sets have been recorded in many different social situations with different degrees of spontaneity and naturalness. The most viable existing candidate was the AMI meeting corpus [50] (discussed in detail in Section 3.1.4 along with other relevant data sets). However, only a portion of this corpus is naturalistic and it was not suitable for the tracking method employed in Chapter 4 and subsequent chapters. A new dataset that was suitable for this study was created. This thesis attempts to use a specific social context because behaviour is expression and perception is dependent on social context. A context was chosen that may have useful applications i.e. it is a commonly experienced situation that is important for interpersonal relationships. Also, the selected social situation should be easily reproducible in order to reduce resources requirements and to enable future work in the same social context. However, database recording and annotation is a challenging task due to the sensitivity of humans to context, practical considerations when recording usable video and audio, and the resource requirements for annotation [66]. A new corpus was recorded based

on informal conversations between two people. This corpus was publicly released to assist further research.

Context is significant in determining the way people perceive NVC [142]. One significant contextual factor is culture [204]; different cultures often have unique meaningful gestures [201]. If a single group of annotators is used and the ratings of annotators combined to form a group specific consensus label, these perceptual cultural differences would be ignored. Using annotators from a single culture results in annotations based on perceptions that are specific to that culture. An alternative would be to use independent sets of expressors and annotations, each set from the same culture. This would allow studies to identify NVC signals specific to that culture. However, the creation of such a dataset is beyond the scope of this thesis which focuses on cultural perception differences. To account for the cultural differences in perception, observers from three different cultures were used to create culturally specific annotations. This can be used as a basis for training and testing automatic recognition systems that are better suited to recognise NVC in a way that is more like a human observer. This is shown by confirming the presence of cultural differences in the annotation data[5] and the automatic system trained on a specific culture is better at recognising NVC signals[6].

Automatic recognition of NVC is difficult because there are many possible ways of expressing a particular NVC. Each person has a particular style of expression [45], as well as being influenced by many contextual factors. During most social situations, there are multiple sources of body and face motion that do not relate to NVC, including lip movement caused by speaking, ambient motion (small body movements that are continuously present, e.g. breathing) and the subset of emotions which are not involved in NVC. As Krauss et al. [170] said "All hand gestures are hand movements, but not all hand movements are gestures[...]". NVC signals are expressed with a duration ranging from a fraction of a second to several hours [5, 322]. NVC also involves visual, audio, tactile and other signals to convey messages [166]. It can be difficult to encode all this information in a form that is suitable for machine learning. This makes the learning of associations between gestures and NVC meaning very challenging. This

[5] see Table 6.3
[6] see Table 7.2

thesis focuses on visual information in the human face. These visual signals are rich in NVC information but considering the face exclusively is far from being a comprehensive view of NVC.

1.3 Main Thesis Contributions

To address these challenges, the main contributions of this thesis are:

1. The design and publication of a new corpus to study human behaviour in an informal social situation. Minimal constraints on participants were imposed to allow spontaneous, natural conversation, which is necessary to gather data that is suitable to study NVC. The corpus is publicly available for further use[7] [287]. This database differs from previously available datasets (discussed in Section 3.1.4).

2. A crowd-sourced method for multi-cultural annotation. This provides a basis for studying automatic NVC recognition from different cultural perceptions. The labels used for annotation were based on NVC message meaning. Cultural differences in NVC perception were found to be present.

3. Design and evaluation of an automatic system for NVC meaning recognition. Various feature extraction methods and classifiers were compared to find an effective approach. NVC recognition was also approached from a regression perspective to produce dimensional, continuous valued label predictions. Effective performance was observed using a system based on tracking facial features, computing distances between pairs of trackers, then forming features based on their statistical properties as the basis for regression. Feature selection was also employed and found to significantly improve performance. Using an automatic system trained on culturally specific annotation data was found to result in a higher performance than using annotations from a different culture.

4. Automatic techniques that were developed as part of the recognition system were applied to automatically identify patterns of human behaviour. The presence of

[7] http://www.ee.surrey.ac.uk/Projects/LILiR/twotalk_corpus/

interpersonal coordination in human behaviour was confirmed and quantified for the face. The feature components that were most relevant for NVC recognition were identified. Analysis was performed using both quantitative methods and by direct visualisation of facial regions involved.

The outcomes of the thesis are:

1. a new, publicly available NVC corpus that has been annotated by people from different cultures,
2. provided additional evidence that there are cultural differences in NVC perception,
3. an effective method for natural NVC recognition and
4. demonstration that culturally specific NVC recognition models leads to higher performance.

1.4 Overview of Thesis

Chapter 2 discusses existing research relating to NVC, classification, machine learning and feature extraction . The recording of the NVC corpus is discussed in Chapter 3. This corpus is then used as a basis for studying automatic classification, as described in Chapter 4. Chapter 5 investigates the effect of interpersonal coordination of behaviour between the two conversation participants, and attempts to use the reaction of a person to infer the NVC of an unseen subject. Collection of annotation data from culturally distinct groups is described in Chapter 6. This annotation data is used in a study of a culturally specialised regression as discussed in Chapter 7. The facial deformation is encoded using a geometry based feature extraction method but this feature contains redundant and irrelevant information. Chapter 8 uses feature selection to isolate a set of features that improves NVC regression performance. General conclusions are drawn in Chapter 9.

To effectively communicate, we must realize that we are all different in the way we perceive the world and use this understanding as a guide to our communication with others.

Tony Robbins

2

Research Context of This Work

One of the central themes of this thesis is Non-Verbal Communication (NVC) and, for clarity, it may be useful to define this term and compare it to the related concept of emotion. NVC is the process of communicating by means other than words [166]. However, Knapp and Hall [166] argued that it is difficult to separate verbal from non-verbal behaviour and to determine if the expression "by means other than words" refers to the way NVC was expressed, or the interpretation that is assigned by an observer. They did say that the broad but inexact definition will serve, as long as these issues are understood and appreciated. For the purposes of this thesis, communication is considered as an intentional [151, 239] action of exchanging thoughts, opinions or information (which contrasts with the view that communication can be expressed with or without intent). This perspective of NVC is similar to Ekman and Friesen's definition of "communicative non-verbal behaviour" which they define as "those acts which are clearly and consciously intended by the sender to transmit a specifiable message to the

receiver" [93]. The exact manner in which an NVC message is communicated can be carefully planned, or performed without conscious awareness [166, 176]. This means that some NVC signals are intentionally expressed while the expresser is unconscious of the exact mechanism by which they are expressed. NVC can also be defined in terms of behavioural preferences of the communicators themselves, which relates to the use of gestures, posture, touching behaviour (haptically), facial expressions, eye behaviour and vocal behaviour [166].

NVC is important for understanding the meaning of communication, because the verbal component is often complemented or augmented by NVC [14, 165]. Non-verbal signals are often used to regulate a conversation, including turn taking [161, 21]. It may be helpful to clarify the terms "vocal" and "verbal". Verbal means to be or pertain to words, while vocal means to be or to pertain to voice. There can therefore be spoken NVC signals, which includes prosody and various non-word utterances such as gasping, groaning, laughing and sighing. The non-verbal component of the voice is also referred to as "paralanguage". The audible part of NVC is not explored in this thesis, with the focus being on visual, and specifically facial NVC signals.

This thesis treats non-verbal communication and non-verbal behaviour as distinct concepts (similar to [170, 134]), although these terms are sometimes used interchangeably [222]. Non-verbal behaviours that convey information, called "informative nonverbal behaviours" by Ekman and Friesen [93], are defined as acts that "elicit similar interpretations among some set of observers". Informative nonverbal behaviour includes both intentional and unintentional transfer of information. A separate issue is if a particular behaviour is communicative, in the sense that the behaviour was intended to be expressed by the sender to convey a particular message. As previously mentioned, Ekman and Friesen defined "communicative nonverbal behaviour" as "those acts which are clearly and consciously intended by the sender to transmit a specifiable message to the receiver" [93]. The terms NVC and communicative non-verbal behaviour are treated interchangeably in this thesis and are characterised by their intentional expression [134, 93, 176] and the use of a shared decoded meaning [134, 93, 176, 332]. Similarly, Burgoon et al. [46] limited the scope of NVC to behaviours that "are typically sent with intent, are used with regularity among members of a social community, are typi-

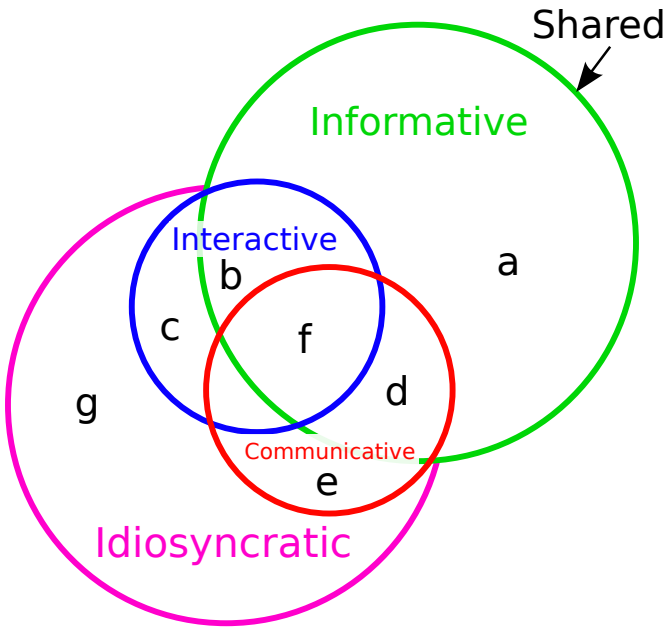

Figure 2.1: A set diagram showing the relationships between various types of non-verbal behaviour, adapted from Ekman and Friesen [93], p. 57. This thesis aims to address communicative non-verbal behaviour, however an annotation's perception of communicative non-verbal behaviour is limited to subsets d and f.

cally interpreted as intentional, and have consensually recognized interpretations". An implication of this definition is that some non-verbal behaviours are not communicative [93, 105]. Ekman and Friesen define a third type of non-verbal behaviour: interactive non-verbal behaviour which "clearly modify or influence the interactive behaviour of the other person(s)" [93]. These definitions of non-verbal behaviour are overlapping concepts; the relationship between them is shown in Figure 2.1.

In contrast to the definition of NVC above, it is popular to define NVC in a much broader sense. This view considers communication to encompass all forms of information transfer, including unintentionally expressed behaviour and informative non-verbal behaviour. The broader definition was described by Fiske [115] as the "semiotics school" and the narrower NVC definition (in the previous paragraph) as the "process school", with both approaches being necessary for appreciation of the field (see also [265] for a broad literature review). Motley argued that, irrespective of the controversy of definitions, intentional and unintentional behaviour are distinct [217]. The broad definition

excludes the possibility of non-communicative non-verbal behaviour because any behaviour can be potentially interpreted, therefore "one cannot not communicate" as claimed by Paul Watzlawick [11]. However, this definition was considered to be overbroad by MacKay [191], leading him to joke that by this broader definition, the study of NVC would "cover every interaction in the universe except the use of words!"

Mental state is a broad term for describing temporary mental conditions that have characteristic properties (see [242] for a review). Mental states include emotion, attitudes, moods and cognitions [198]. Many mental states manifest themselves as an outward expression and this expression can be intentionally modified to form a partly impulsive and partly controlled expression [120]. This means mental state displays have both voluntary and involuntary aspects. However, some mental states are not necessarily externally observable, which distinguishes them from NVC which is always externally observable. Emotions are a group of mental states that have connections to particular behaviours, have particular physiological manifestations (such as facial expression) and are also subjective experiences [119]. They tend to be triggered by stimuli, which are evaluated by a person, and spontaneously results in an emotion. Emotions are often expressed by facial expressions, vocal utterances, behavioural changes, and physical responses. The facial area is particularly important in emotion perception [98]. A facial gesture is a motion executed with the facial muscles and may be associated with mental states, emotion or NVC. Emotion is an ill defined concept [31, 286, 112]. However, this does not imply that there is no such thing as emotion, nor that it is not a subject worthy of investigation.

NVC is a communication act that is only expressed in social situations. While some hold the view that emotions are also limited to social situations, others have observed that some forms of emotion can be expressed in non-social situations such as reading [195] or dreaming [225]. In a similar way to NVC which typically lasting from a few seconds [175] up to hours or longer, emotions can last from a fraction of a second (e.g. surprise) to hours or even longer (such as with empathy) [5, 322]. The choice of clothing is an example of NVC that has a long duration. A person's internal emotional state is sometimes manifested by facial expression, but the extent to which this is a direct relationship is controversial [22]. At one stage Ekman claimed that six emotions (Anger,

Disgust, Fear, Happiness, Sadness and Surprise) occurred across cultures [96] and called them "basic emotions", but later appended additional emotions to this list [97]. He holds the position that different emotions are physiologically discrete and separate. This contrasts with approaches that describe emotions using continuous value labels [66].

A social signal is 'a communicative or informative signal which, either directly or indirectly, provides information about "social facts", that is, about social interactions, social attitudes, social relations and social emotions' [250]. Social signals include 'interest, determination, friendliness, boredom, and other "attitudes" toward a social situation' [240]. The term social signals is used in animal behaviour[181] and for human non-verbal behaviour [240]. The definition does not specify if the signals are necessarily verbal or non-verbal and many authors use the term "non-verbal social signal" for clarity [230, 324]. They are distinct from "social behaviours" which "convey information about feelings, mental state, personality, and other traits of people" and includes expressions of politeness and empathy [324]. The next section discusses the factors on which NVC expression depends.

2.1 What Factors Influence NVC Expression?

There are many factors that influence human expression and perception, and specifically NVC. If an automatic system is intended to be trained and deployed in a single environment, this will be of little concern. Although encoding based on motion is not sensitive to context, a system that is to recognise communicative intention for multiple people, or in multiple social and cultural situations, needs to account for contextual factors that can change how NVC messages are interpreted[1]. NVC is largely determined by the social situation in which it is used and therefore it is important to study NVC in natural social situations [28]. Social context is also a factor that is used by human observers to interpret the behaviour of other humans, and humans are not reliable

[1] "From this example, it is obvious that in order to determine the communicative intention conveyed by an observed behavioural cue, one must know the context in which the observed signal has been displayed" [237]

observers when this context is removed [142]. Social context, also referred to as social environment, "encompass[es] the immediate physical surroundings, social relationships, and cultural milieus within which defined groups of people function and interact" [25]. A waving gesture can be a greeting or a sign of distress depending on the context in which it occurs. Although cultural differences in expression exist for many NVC signals, some signals have a similar appearance across cultures and social situations. There has been little research of the automatic recognition of NVC messages that are specific to contexts, with most existing approaches only considering a single context or seek to generalise NVC across different contexts.

Factors which influence the expression and interpretation of human behaviour include:

- culture (this is discussed in depth in Section 6.1),
- gender,
- personal style,
- personality and
- social situation.

People naturally vary in expressiveness; some individuals being animated while others being comparatively inexpressive [7]. Buck claimed that encoding and perceiving accuracy for NVC was dependent on gender, personality and age [45]. However, these studies assume that there can be "correct" and "incorrect" interpretations of emotion and NVC. This view seems questionable because the use of posed behaviour does not necessarily imply that the samples are directly associated with an objective, exemplar basis of human behaviour. However, the study does highlight the differences in the interpretation. Gender difference in expression style was investigated by Morrison et al. [216], who found there were specific facial movement styles that could be used by humans to identify gender. Some medical conditions, such as schizophrenia spectrum disorder, can change how NVC is expressed [43]. All these studies provide evidence that there are individual variations in how NVC or emotion is expressed.

When humans experience emotion, the emotion often manifests itself in body language and facial expression. Emotion expression is based on many factors. The mapping from emotional state to emotional expression can be thought of as a set of rules, according to Ekman and Friesen [94] with each culture having a specific set of encoding rules. Other researchers have extended this idea to specific social situations having distinct display rules [113, 42] that encode the internal state. For this reason, a person expressing an emotion is sometimes referred to as an "encoder". Research related to the effect of culture is described in more depth in Section 6.1. Studying the mapping from emotion to expression across cultures can be challenging. To demonstrate that cultural display rule differences or similarities exist, the underlying emotions used must be shown to be equivalent across cultures. NVC expression may also have encoding rules which are analogous to encoding and display rules. Just as emotions have shared social norms and expectations that are used when managing and modifying emotional displays in different social circumstances, NVC expression is also dependent on social circumstances.

Social situation is also a factor in NVC and emotion expression. The effect of social situation itself shows cultural, gender and personality differences [17] making these factors interdependent. Social situation is significant for capturing a corpus of NVC behaviours, because the situation in which the recording takes place affects the type and frequency of observed behaviours. It is often convenient to record posed data, because behaviours directly correspond to pre-determined labels and little time is wasted on recording uninteresting behaviour[2]. All human behaviour occurs in a situational context, although acted behaviour may be a special case in that it has a context in terms of social interaction with the audience. Unfortunately, human behaviour is significantly different in posed situations when compared to spontaneous behaviour. Cowie [66] argues that the use of posed data cannot be completely excluded but posed data should not be used uncritically. There has been a recent shift in the automatic human behaviour recognition community to use natural data, rather than posed. However there is a wide range of approaches to collect so called "natural data" to the point that the word can be misleading. The definition of natural language proposed by Stubbs

[2]Even if acted data is considered as having no social context, the absence of a social context is a factor in NVC expression.

[295] can be adapted to suit NVC: that natural NVC occurs "without any intervention from the linguist [or experimenter]" and "is spontaneous in the sense of unplanned, and which is composed in real time in response to immediate situational demands". Applying this to NVC, this definition excludes posed examples of NVC, as well as NVC based on artificial tasks, stages scenarios, role play tasks or experimenter controlled stimuli (such as a "Wizard of Oz" apparatus). The issue of social situation is discussed in more depth in Section 3.1.1. Just as the expression of NVC depends on many factors, the perception of NVC is also dependent on multiple factors. These will be discussed in the next section.

2.2 Perception of NVC

The interpretation of emotion is based on various contextual factors, such as culture, social situation, age, etc. Building on the idea of encoding rules, Argyle [17] suggested there exists a mapping from observable behaviour to a meaningful interpretation and termed this mapping as "decoding rules". Several studies have found context and conditions that affect perception of human behaviour, including NVC and emotion. Terracciano et al. [305] found there is a relationship between personality and emotion perception. Matsumoto et al. [203] claimed that accuracy in emotion recognition was correlated with personality measures such as Openness and Conscientiousness. Personality differences of experimental participants, such as neuroticism, were associated with different subject gaze patterns being observed during an emotion recognition task [241]. However, little work has been conducted in the association between personality differences and NVC perception. Behaviour mirroring, which is the tendency of people to adopt the behaviour of another person in a conversation, is more pronounced in people who had tested highly on an empathic personality measure [52]. Lanzetta and Kleck [173] found that a person's NVC perception ability was related to personality and that highly able subjects were themselves difficult subjects for others to read. All these findings provide evidence that both personality and individual style have a role in the perception of emotions and NVC.

Another factor in behaviour perception is gender. Vassallo et al. [320] found that

gaze patterns during evaluation of emotion were different between genders and that females arrived at a judgement faster than males. Gender based differences in annotator judgements have also been observed in several studies [45, 305, 6]. Context can be a factor in perception, but this may still be specific to certain cultures and personalities. The way face images are viewed can affect perception. The presence and expression of surrounding faces changes the judgement of a central face in some cultures but not others [200]. Goren and Wilson [126] found that intense, posed emotion is easier to categorise than weak, posed emotion, as well as finding that emotions are harder to rate accurately if viewed by peripheral vision. Perception of emotion was significantly affected by the level of familiarity with the person being observed [142]. El Kaliouby et al. [101] found that while basic emotions (Ekman 6 and contempt) were almost instantaneously recognized by humans, temporal context improved human recognition for complex emotions, such as interest, boredom, confusion, etc. All these factors make annotation a challenging task, because these variables cannot be easily controlled while maintaining the naturalness of the data. Using multiple annotators and finding a consensus is one common approach. However, personality, relationship and familiarity of people in a social situation are important contextual factors in the expression of NVC and using a consensus score discards this aspect of context.

2.3 Supervised Learning for NVC Recognition in Video Sequences

One of the aims of this thesis is to create an automatic system to perform facial, intentional NVC recognition, based on previously observed examples of behaviour. As well as directly relevant studies in the field of automatic behaviour recognition, this review discusses research from other fields if they are particularly relevant to this thesis. Video recording of NVC has been previously been used in the behavioural sciences [202], as well as studies into facial biometrics [127], automatic lip reading [232], character animation [34] and affective computing systems [237, 26, 344].

Based on video of facial behaviour, training data and manually annotated labels are used to create a model; this process is known as supervised learning. Based on the

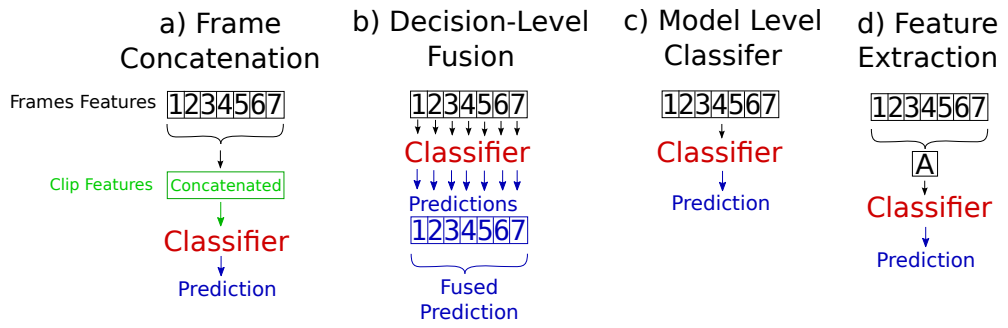

Figure 2.2: The common approaches to using a classifier with video sequences. Feature extraction and machine learning have a different role in each case.

model, labels for unseen samples can be automatically predicted. The input data and labels are digitised to allow computers to process the input data. Supervised learning is often divided into feature extraction , which provides a set of higher level features, followed by a classification technique. Given an ideal training set with total coverage of the feature space, the use of feature extraction and a sophisticated classifier would not be necessary; nearest neighbour classification [64] would be sufficient. However, the available data sets contain a limited number of samples and also contain noise. To improve the performance of a classifier on limited training data, visual changes that are not relevant to the task at hand should be separated or removed from the input data. This improves the robustness of an automatic system and is usually achieved by manual design of the system or by feature selection. This is often accomplished by feature extraction , which aims to improve robustness and reduce the quantity of training data required. Almost every application of supervised learning uses feature extraction which transforms the data into a "higher level" representation; common approaches are described in Section 2.4. The feature extraction technique used in an automatic system is usually selected manually. There are also many supervised classification methods (described in Section 2.5). The type of classifier used may dictate which feature extraction approach is optimal and visa-versa, so these issues are closely interrelated.

Once feature extraction has been performed on a sequence of video frames, the variable length of examples and temporal nature of a video must be considered to achieve effective automatic recognition performance. This section only considers the processing

of visual information and does not consider the issues of fusing audio and video, which is beyond the scope of this thesis. A temporal NVC signal can be recognized using data from multiple observations by using feature extraction which encodes temporal variations early in the recognition process. Alternatively, a model based recognition system may be used, which deals with temporal variation at a higher level. Feature extraction of multiple observations that were acquired at different times is referred to as "temporal fusion" [319, 167]. These feature extraction approaches will now be described in more detail, in the context of temporal, uni-modal video based recognition (see also Figure 2.2).

- Sensor level feature extraction consolidates raw video image images into a combined raw feature. This is rarely used in the context of human behaviour recognition.

- In one form of temporal feature extraction, all frames can be concatenated and directly used by a standard classifier. This can be directly applied to fixed length examples. (Figure 2.2 a) Frame concatenation is a special case of feature-level encoding. If different modalities are sampled at different frequencies or if samples are of varying length, re-sampling can enable feature concatenation.

- Decision level fusion: Each frame can be classified individually using standard machine learning, then the predictions combined and used in a second classifier step to provide an overall overall clip prediction. Decision level fusion may use multiple samples from one or more types of sensor [311]. (Figure 2.2 b). Variants of decision level fusion include rank level fusion and match score fusion. Rank level fusion: multiple recognisers rank possible hypotheses and these rankings are combined to form a final label. Match score fusion: prediction labels are provided by multiple recognizers based on multiple, individual observations and these labels are then combined to generate information for decision making.

- Model level recognition: individually encoded frames can be directly used by some types of machine learning methods. The order of the frames in the clip can be used (as typically done with Hidden Markov Model (HMM), Conditional

Random Field (CRF)) or ignored (e.g. as Nearest Neighbour or Multiple Instance Learning (MIL)). (Figure 2.2 c)

- Feature extraction methods encode how low level observations vary over time in a way that can be interpreted by a classifier. Frames can be summarised or combined into a fixed length vector, then classified using a standard supervised learning technique. This approach is used in Sections 4.7 and 7.3.1. (Figure 2.2 d). One simple approach is to concatenate frames into a single vector (Figure 2.2 a).

Concatenating frame based shape or appearance features before classification is rarely used because videos of varying lengths cannot be directly compared and the approach tends to be sensitive to the speed of activity in the video. For varying length videos, one approach is to re-sample the frame features to produce a fixed length vector, as done by Pfister et al. [249]. However, the feature vector can be sensitive to whether an event of interest occurs at the start of a clip, or at the end; this may not be desirable if the occurrence of the event at any time is of significance. This is not an issue for approaches that use unsegmented videos [228].

Feature extraction is a popular group of approaches that combine one or more frames into a feature vector, usually of fixed length. Multimodal feature extraction by up-sampling video features and concatenating them with audio features was performed by Petridis and Pantic [246, 244, 245] and others [252, 207, 224]. This can increase the dimensionality of the data, which may reduce the performance of the final system. Other feature-level extraction approaches try to encode temporal variations [79]. Valstar et al. [317] calculated the mean and maximum change in the displacement and velocity of feature points during facial deformations, which is a form of temporal feature extraction at the feature level. This reduces the dimensionality of the data while hopefully retaining temporal information about the face. In a similar way, face deformation based on heuristic features was encoded by taking the variance of each feature by Datcu and Rothkrantz [72]. As well as using the mean and standard deviation of features in a clip, Petridis and Pantic [246] fitted a quadratic curve to each feature component to encode temporal variation. Most of these approaches use geometric deformation based

features, however temporal encoding has also been used for texture and audio. Aligned facial textures were compared using the mean shape and appearance features by Fang and Costen [109], however, using only the mean does not encode temporal variation information. For audio features, Grimm et al. [131] used seven simple statistical measures on the features, as well as the feature's first and second derivatives, which may be useful in encoding the rate of feature variation.

Fusion after matching combines the predictions from multiple classifiers to form a final decision. This is also known as match score fusion and is a form of decision level fusion. This may be used to combine different types of features, different modalities, predictions from independent video frames or the predictions of different classification methods. The most popular decision level fusion methods are the "sum rule" and rank based voting [164]. The sum rule takes the class dependent average of the predicted probabilities and selects the class label with the highest score. Rank based fusion has each classifier rank each label and these are combined, using one of several approaches, to form a final prediction. One ranking fusion method is majority voting, which takes the highest probability label from the classifiers and the label with the highest proportion is taken as the fused prediction. Audio and visual decision level fusion has been used in many studies [311, 251, 246, 243, 158]. Pfister et al. [248] used majority voting to combine predictions of three types of classifier for visual features. Three different rank based fusion methods were compared by Akakın and Sankur [10] to combine different types of visual features, although there were no appreciable performance differences between the ranking approaches. Audio and video modalities were fused at the decision-level by the sum rule by Petridis and Pantic [246]. The sum rule allows the decisions to be weighted to change the emphasis on different modalities, as done manually by Kanluan et al. [158], or automatically by Logistical Linear Regression (LLR) fusion as performed by Lucey et al. [189]. Oikonomopoulos et al. [228] used voting to combine information from different human action models. Ko et al. [167] showed that dynamic time warping can result in a higher performance than a HMM approach for recognizing hand gestures based on multi-sensor fused data. Based on these studies, combining multiple modalities with decision-level fusion often results in a large increase in performance compared to a single modality (although exceptions exist to

this pattern [224]) and often has recognition performance advantages over feature level fusion (although some studies have shown that the performance using feature fusion is comparable to other approaches in some cases [207, 246]).

Rather than combining features or decisions, feature vectors from multiple frames may be considered directly by a machine learning algorithm. There are two general classes of algorithms: sequential temporal model classifiers (e.g. HMM, CRF) that consider an ordered group of items, and "unordered set" classifiers (e.g. MIL) which considers a set of items in which the ordering is not considered. These are among the most popular approaches in the recognition of human behaviour. They are discussed in more depth in Section 2.5.

Petridis et al. [243] used a different approach to combining multimodal information by using the audio mode to predict the expected visual features and visa-versa. A positive prediction is made if the predicted features match the observed features and this approach exceeded feature level fusion performance for laughter vs. speech discrimination.

2.4 Encoding Facial Information

Facial features encode shape information by shape or appearance changes, or by combining both types of information. The common approaches will now be described and some of their strengths and weaknesses will be discussed.

2.4.1 Encoding Behaviour Using Appearance Based Features

Changes in facial expression often cause changes in facial appearance due to wrinkles, deformations that change the visibility of the inside of the mouth, etc. However, to make effective use of limited training data, images are typically aligned to a canonical head position or to a canonical head shape. This ensures correspondence between local areas of different face images is maintained and the effect of head pose changes and translation is reduced, making the method much more effective. An alternative is to use changes in an overall image without using alignment which would be effective in

encoding head movements, as done by Escalera et al. [106], but this cannot accurately encode subtle facial deformations. Some papers on emotion recognition deliberately ignore the image alignment problem and instead focus on the robustness of later steps in the automatic process [285].

Some approaches to face alignment begin with interest point detection; the face can be subsequently re-scaled or affine transformed to a canonical alignment [345]. Yang et al. [337] used SIFT interest points and a similarity transform. Yang and Bhanu [339] found point correspondences between images using the SIFT-flow technique and removed shape information in the alignment process. This would help to reduce the effect of identity in emotion recognition. Zhu et al. used Lucas-Kanade based local correspondences to perform a non-rigid transform [345].

Other approaches for face alignment use whole face detection or a more general model for the overall face shape or colour. The most popular basis for face alignment is the Viola and Jones object detector [325], with several studies using this as a basis for image alignment [158, 27, 212]. Skin colour based detection was used by Feng et al. [114] to align images; this was robust to illumination changes and, to a limited extent, to pose changes. Some models attempt to fit an expression model (which is often inspired or validated using Facial Action Coding System (FACS) [63, 9]) or a shape and appearance model to the observed face, or be limited to variations of shape, head position and pose. Shape based approaches will be discussed below in more depth but there are a few studies that use face models for the purposes of alignment pre-processing before extracting appearance based features. Pfister et al. [249] used an active shape model to achieve alignment. Active Appearance Model (AAM)s may also be used for normalisation [172] but it common to use the model's parameterisation directly for recognition, rather than to the normalised face as a preprocessing step. Dornaika and Davoine [84] used a 3D head model to not only normalise the face translation and pose, but also to remove shape information from the features. Other whole face based approaches avoid shape models and use image information more directly to determine an appropriate alignment transform. Visser et al. [326] used Principal Component Analysis (PCA) on training images to perform lip localisation, however this was computationally expensive. The alignment transform may be directly found by the affine transformation of an

input image on to a canonical image, by maximising pixel intensity correlation [310]. Rudovic and Pantic [266] used a shape constrained Gaussian process to fit a facial point based model while retaining an anatomically plausible shape. Dhall [78] et al. used Constrained Local Model (CLM) to normalise head positions for emotion recognition, which fits a parameterised model to landmarked positions based on an ensemble of local feature detectors.

Image alignment is intended to remove appearance variations caused by face translation and pose. Often a simple affine or re-scaling transformation is employed. However, natural conversations contain extreme head poses. Until recently, alignment of images based on detection of facial features over a wide range of head poses was an unsolved problem [346]. An affine transform from an input image to a canonical face is challenging because of self-occlusions [339]. For these reasons, using an approach that depends on face alignment to frontal view may fail at extreme poses.

Once the input image has been aligned, appearance features may be extracted that correspond across multiple faces. Two broad approaches are employed: holistically encoding pixel information or part based methods. These types of appearance based features will be discussed.

- Holistic approaches use dimensionality reduction techniques to encode changes in overall pixel intensity. Visser et al. [326] and Seguier et al. [282] performed PCA on the lip region of interest. Chew et al. [54] used shape normalised textures to recognise Action Unit (AU)s and compared this method to local texture approaches under the effect of noise, on four public datasets. Texture descriptors may be applied to every pixel in the image as the basis for classification [27, 212]. If a transform is applied on a per pixel basis, the feature vector may become excessively large. Using histograms of texture reduces the representation to a manageable size and also reduces spatial sensitivity, which may be advantageous in improving robustness to insignificant differences in face shape. Kanluan et al. [158] performed Discrete Cosine Transform (DCT) on lip and eye pixels. He et al. [139] used Independent Component Analysis (ICA) on facial images to find independent modes of variation. These methods are useful for finding low frequency

spatial information, which is often more relevant than using individual pixels or high frequency components, which correspond to very small areas or very small changes that are unlikely to be significant on their own. However, changes in individual components of these methods typically correspond to a global change in the region of interest and would involve multiple muscle movements. It is likely that features that do not isolate changes based on individual facial muscles would be sub-optimal.

- Rather than encode an overall image, a feature can attempt to encode local texture information near points of interest. This makes localised changes in the face affect only some feature components. As mentioned, this decoupling of the different parts of the face may be useful because it is expected that local changes in the face are useful for recognition. Texture descriptors may be applied to limited regions and encoded using histograms, which provides a more compact representation [114, 285, 249, 213, 339, 335]. There are several popular texture descriptors used in facial recognition, particularly Gabor filters ([27, 328, 263]) and LBP features ([114, 285, 339, 213]). Recent work has considered layers of texture descriptors. Senechal et al. [283] used Local Gabor Binary Pattern (LGBP) histograms for AU recognition, which are LBP operators applied to Gabor filtered images. Tingfan [335] compared using a single layer (Gabor energy filters and LBPs) with double layer texture filtering, finding that double layer is more effective in multiple data sets. LBPs were extended to encode temporal variation, and named LBP-Three Orthogonal Planes (TOP), by Pfister et al. [249], however considering changes on a short time scale, such as consecutive frames, may not be optimal for all facial gestures. LBP features are also noted for being relatively robust to illumination changes, and in some forms are also robust to rotation. Other approaches include edge based texture descriptors, such as Moore and Bowden [212] using Canny edge detection to form chamfer images as the feature extraction step. Yang and Bhanu [339] used Local Phase Quantization (LPQ) features that exceeded LBPs performance for emotion recognition on acted data. Jiang [154] extended the LPQ to consider temporal variation (LPQ-TOP) and found it exceeded performance of LBP-TOP.

Donato et al. [83] compared many appearance based approaches and found Gabor features with ICA dimensionality reduction was the most effective for facial expression recognition.

As previously mentioned, images are usually aligned before appearance features are extracted. While these features have been successfully employed for many constrained datasets, appearance based features are only as good as the face alignment process, which may be problematic for extreme head pose. Appearance based features encode information about the face that can be used as the basis of classification, but also includes information that is not necessarily relevant. Depending on the behaviour under consideration, these irrelevant appearance differences include facial hair, wrinkles, skin colour, glasses, etc. Although some emotions are associated with colour change in the face, colour information is rarely considered because differences in skin colour make interpersonal facial colour comparisons problematic. Various approaches can be employed to remove personal differences in features (Section 2.4.5). Appearance features also may be non-optimal if significant facial movements occur that do not cause a detectable change in local appearance. Facial deformations are expected to be important for automatic NVC recognition. For this reason, many approaches attempt to encode face shape directly. The next section will consider these types of features.

2.4.2 Encoding Behaviour Using Shape Based Features

Shape based features are used to encode information about facial deformation and head pose. Because many human NVC signals are based on body and face gestures, encoding the shape captures this type of information. Shape based features usually involve fitting a model to an observation image. The body part selected for feature extraction and sophistication of the model depend on the task that is being attempted. Various approaches have used simple head models that can encode head pose changes, or both the pose and facial expression. Yang et al. [340] used a simple 2D model to encode the face. Chen and Ji [53] apply Dynamic Bayesian Networks to combine facial tracking with expression recognition in a hierarchical framework. Zhu and Ramanan [346] used mixtures of trees with a shared pool of parts to localise faces, estimate pose,

as well as locate points of interest in unconstrained photographs. Petridis et al. [244] used a 3D cylindrical model to track the face, with six degrees of freedom corresponding to translation and rotation. This would not encode facial expression but should be sufficient for gestures based on overall head movement. Dornaika and Davoine [84] used a more sophisticated model including deformation to fit facial expression. Zeng et al. [343] used a 3D model based visual tracker for expression recognition. However, complex models are computationally demanding, have higher training requirements and usually require advanced methods to fit the model reliably and robustly. Other approaches use 3D shape based on structured light capture systems for emotion recognition [273, 108, 53]. The face is of most interest for NVC recognition, but may be complemented by other parts of the body. Occasionally other parts of the body are used for model fitting and recognition, including the shoulders [244] or the overall body pose [32].

Some studies use a hybrid approach to extract a model, such as AAMs, and then use only the components corresponding to the shape for recognition [123]. This has been done by heuristic feature extraction from AAM features [72], as well as distance ratios of points of interest [301]. Other approaches use both shape and appearance, which will be discussed in the next section.

Another approach is to treat the body as independently moving parts, localise the position of each part in the frames of the video sequence and use this information directly, or fit a model based on the independent parts. This is conceptually simpler than fitting a complex model with inter-dependent parts. However, position and motion of areas of the body can provide information for the movement of nearby body areas and this is not used by tracking independent points of interest. Tracking attempts to localise a feature of interest in a series of video frames. Tracking uses the assumption that a point of interest has a locally constant appearance. This separates motion from other non-shape changes in the video. However, large changes in appearance caused by pose changes or occlusions make tracking spontaneous videos difficult. Lucas and Kanade [188] proposed a method based on minimising the least squares difference of a training patch to a test image by gradient descent. However, it can suffer from local minima, noise and occlusions which can lead to tracker drift. An alternative is to use a particle filter based approach, which is robust to noise by using multiple

hypotheses of tracker position [148]. However, this approach does not scale well to high dimensionality. This was addressed by partitioning the model space components into groups by Patras and Pantic [238]. This tracking method is limited to small head pose changes of approximately 20 degrees [316]. This is because of the change of appearance caused by head rotation breaks the assumption of an unchanging local appearance near to a tracked point of interest. Chew et al. [54] used a CLM to detect AUs, which uses detections of facial points of interest as the basis to fit a shape model. CLMs [70] are similar to AAMs but generate likely feature templates that correlate with the target image, rather than trying directly fitting a model to pixel intensities. Baltrusaitis et al. [24] extended CLMs to 3D and applied it to head pose tracking. Liwicki et al. [185, 184] proposed a kernel PCA method, which retains desirable properties of PCA while being robust to outliers; this encodes behaviour as shape information. They applied the method to different computer vision problems, including visual tracking and found it was better than 4 other state-of-the-art tracking methods in most video sequences.

The approach used in this thesis is the pre-existing tracking technique proposed by Ong et al. [231] called Linear Predictor (LP) tracking; this method encodes behaviour as face shape changes. This method learns a linear mapping from intensity changes in a sparse template to a tracker positional offset. The tracker is trained on multiple training frames which improves its robustness to appearance change, including an amount of head rotation. The maximum tolerated head rotation is not known, but it is suitable for tracking spontaneous human behaviour. The tracker does not automatically recover from occlusions, so the tracker must be manually re-initialised when a feature becomes visible. This could be addressed by incorporating an automatic feature detector to re-initialise tracking positions. This tracking process results in the facial behaviour being encoded as a series of shape deformation frames. The technical details of LP tracking are discussed in Appendix B.

Tracking data encodes facial deformation changes and can be used directly for behaviour recognition. Tracking is not effective in areas which contain relatively little local texture, such as can be seen in puffed cheeks. Tracking has often been applied to estimate optical flow and this approach encodes the overall face deformation. This

method was used to recognise emotions by Rosenblum et al. [264]. Overall head translation information is usually unrelated to facial expression and is often separated from facial deformation information. However, differences in face shape can prevent direct comparison of optical flow features generated on two different people. Also, optical flow based approaches tend to be sensitive to head rotations and require a constant frontal view of the face. Other approaches use tracking of points that corresponds to known positions on the face. This correspondence makes inter-person comparison of deformations easier. Tracker movements are caused by both head movement and expression changes, but this is not ideal for existing machine learning techniques. Head motion was separated from deformation changes using PCA by Petridis and Pantic [246]. Another way to improve person independent recognition is to remove the effect of face shape by mean shape subtraction [304].

2.4.3 Encoding Behaviour Using Both Shape and Appearance

The previous sections have discussed features that encode either the shape or the appearance of human NVC and emotion. To guide future work, it may be useful to know which approach is most effective, or if these approaches can be combined effectively. Some papers compare shape features to appearance feature approaches and this provides insight into the best approach for encoding behaviour. However, comparisons of specific techniques do not rule out the creation of superior feature extraction techniques. Lien [179] compared three methods (feature point tracking, dense flow and texture descriptors) for emotion recognition of posed examples and found dense optical flow is most effective. Both Fanelli et al. [107] and Akakın and Sankur [10] had similar findings, with optical flow or tracking being effective but optical flow or tracking combined with appearance and shape demonstrating a higher performance. However, optical flow techniques are susceptible to head rotation, which commonly occurs in spontaneous behaviour. Some model based methods extract both shape and appearance information which can be used for recognition. A popular approach for facial analysis is an Active Appearance Model (AAM) [88]. This has been used for emotion recognition and studies have found that while shape is more important than appearance, use of combined shape and appearance has a higher performance [18, 189]. However,

AAMs have difficulty fitting to faces under a wide range of poses and expressions, and require additional specialised models to account for this variability [247, 174]. This increases the training requirements of AAMs, making their application to new subjects very time consuming. Koelstra et al. [169] used motion both history images (which encodes appearance information) and a motion field (which encodes face shape changes) to form local histograms, as the basis for classification of facial actions with motion deformation features being the more effective approach.

The FERA2011 Challenge on emotion and AUs attempted to benchmark different approaches to provide a clear ranking for current approaches [314]. Despite the apparent dominance of shape based features in other comparisons, the best performing approach in this challenge used appearance features on aligned faces [339]. Several other approaches used fusion of both appearance and shape [315, 313]. Senechal et al. [283] used an AAM to recognise AUs but this was found to be less effective than LGBP histograms. Tariq et al. [303] fused local optical flow, Scale-invariant Feature Transform (SIFT) histograms and Hierarchical Gaussianization achieved the best person-specific emotion recognition performance.

Only a single paper used shape feature extraction exclusively but did not have a competitive performance. All the published comparisons of different classes of features have used posed data. The relative importance of shape and appearance is still under debate for this problem, but combining both types of features usually results in improved performance.

2.4.4 Encoding Behaviour Using Audio Features

Paralangauage, which is a form of non-verbal audible communication, provides information that is not necessarily available if only visual facial behaviour is considered. Many studies have employed audible signals exclusively or combined both audio and visual signals (see [281] for a review). Audio feature extraction methods include fundamental frequency (F0) [182, 131], intensity, Mel Frequency Cepstral Coefficients [131], distribution energy in spectrum [308], speech rate [308] and manually extracted features e.g. spectral tilt [182]. Automatic systems have been created that recognise arousal

and valence labelled data [308], categorical emotional labels [182, 136] and affective states [292]. Hassan and Damper [136] showed that finding an appropriate subset of features, using Sequential Forward Selection (SFS), can enable a simple classifier (k-nearest neighbour) can provide a better performance than SVM classification without feature selection. Speaker normalisation is sometimes used to improve generalisation to new subjects [292] or languages not contained in the training data [292, 137].

Many studies use both visual and audio features to recognise behaviour and emotion [245, 158, 135]. There is a close association between facial behaviour and the non-verbal channel of communication [47]. The next section describes further processing that can be applied to features, in an attempt to improve robustness.

2.4.5 Post-processing of Facial Features

As discussed in the previous section, feature extraction attempts to extract relevant information from raw observations while ignoring variations that are irrelevant for the intended task. Many feature extraction approaches consider individual frames and do not consider the overall video sequence. However, facial expressions can evolve over many frames, rather than being restricted to one or two frames. Using the information for multiple frames can be useful to generate better features and to make the final system more robust. Face shape is associated with a person's identity but this is not of interest for automatic NVC recognition. Also, individual feature components can be processed by dimensionality reduction algorithms to attempt to isolate the information of interest. This makes the features more suitable for machine learning, particularly when combined with feature selection.

Using the set of feature values over a long time period, various normalisations may be applied. Face shape for neutral expression cannot easily be inferred from a single frame of video because of the possible presence of expression, but over many frames the neutral face shape is relatively easy to determine, particularly in naturalistic behaviour which often contains long periods of neutral face expressions. The effect of face shape can then be subtracted from the features which improves a system's robustness to identity. Other inter-personal differences, such as facial hair, skin colour and glasses can affect

appearance based features but can be managed in a similar way to face shape. There have been several different feature normalisation methods proposed including mean feature subtraction [152, 304] and scaling features to a standard variance [340] (also referred to as "whitening"). Six audio feature normalisation methods were compared by Wöllmer et al. [333] who found that normalising features to a range of -1 to +1 was advantageous in some cases and was thought to remove identity based differences. Such normalisations are suitable for recorded data but cannot be immediately used for prediction based on a previously unseen face.

Features often encode both information of interest as well as irrelevant information. It can be beneficial to use dimensionality reduction in an attempt to separate relevant information into specific feature components in an unsupervised way. Lien et al. [179] used PCA on dense optical flow features of the face from multiple frames to produce "eigenflows". Fang and Costen [109] used PCA to encode facial feature position motions and Akakın and Sankur [10] applied ICA, Non-negative Matrix Factorization (NMF) and DCT on sequences of facial feature positions. Individual components of features in the eigenvector based methods correspond to deformations of part or all of the face. These deformations are found in an unsupervised fashion and are not necessarily optimal for recognition. Also, the order of events that are encoded in the sequences are either not encoded at all, or encoded in a way that is not appropriate for machine learning.

Another important feature processing stage is feature selection, which is discussed in Chapter 8. The next section describes techniques for supervised classification based on features.

2.5 Classification

One of the aims of this thesis is to create an effective automatic system for NVC recognition. After feature extraction and processing, the final step in an automatic system is often classification, which is the process of automatic prediction of a label based on a test sample observation. Many different classifiers have been applied to

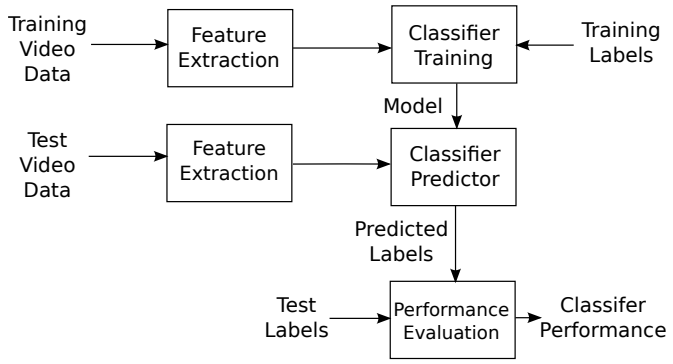

Figure 2.3: Supervised learning uses labelled training data to create a model that can make label predictions for unseen data.

emotion, speech reading and gesture problems. This section provides an overview of the significant existing approaches to classification.

A supervised classifier is used to create a model based on labelled training data. This model can be used to predict labels for unseen test samples. The samples' features are usually the result of a feature extraction process from raw observations such as video frames. In this case, a test or training sample is a video clip of one or more frames. Given a set of training and test data, the performance of a classifier can be evaluated by comparing its predictions for unseen samples to the test data labels, also known as "ground truth". An illustration of this process is shown in Figure 2.3. There is a vast number of classification techniques in the published literature. Specific disciplines often have a set of preferred techniques that have been found to be effective. This is because each classifier has different requirements and makes different assumptions about the problem's characteristics. Also, each feature extraction technique has different properties, so features vary in suitability based on the problem and the specific classifier used. Rosenblum suggested that expressions with greater motion are easier to classify [264]. This section will focus on the use of classifiers in various facial classification applications and will discuss some of the classifier's properties in each context.

There are a few major families of classifiers, each requiring different formats for the input features and input labels. For this reason, it is not possible to apply every classification technique to every problem. The main families of classification techniques will now be discussed, along with the problems to which they have been applied.

2.5.1 Single Instance Classifiers

Single instance classifiers operate on samples that are represented by a single feature vector and a label value. This reliance on a fixed length feature vector makes classification of variable length clips problematic, but this can be overcome by feature extraction or decision level fusion. An early classifier was based on finding the k nearest neighbours for a test sample [64]. This approach is conceptually simple but it has high computer memory requirements. It also considers all feature components as equally significant, which can be problematic if some of the components of the feature vector are irrelevant. This may be the case for NVC and emotion recognition, because typically only part of the face is relevant in determining the label. Donato et al. [83] applied the nearest neighbour technique to facial expression classification. Nearest neighbour is appropriate for problems that have a large quantity of training data. Most classification techniques use a statistical model to approximate the decision boundary between different classes in feature space. One of the earliest classification techniques is Fisher Linear Discriminant (FLD), which attempts to use a linear manifold to separate two classes. Linear discrimination was applied to facial expression by Rose [263]. Unfortunately many problems in facial analysis require the use of non-linear decision boundaries to achieve acceptable performance.

A neural network is a machine learning technique based on an interconnected group of artificial neurons. Neural networks have high training requirements, and high computational and memory requirements. They also generate an internal model that is hard for humans to interpret and therefore only provide limited scientific insight. Their partitioning of feature space is also hard to grasp intuitively. However, neural networks have been used in a wide range of applications including speech reading [326, 282], emotion recognition [264, 244, 334], classifying laughter or no laughter [243] and many other non-facial analysis applications.

Boosting classifiers are a family of machine learning techniques that combine a set of weak learners into a single strong classifier. This is usually done iteratively and greedily, with weak learners being weighted and added to a bank of weighted weak learners. Many boosting approaches limited the weak learners to produce Boolean feature vectors

and also require Boolean class labels. This makes some boosting methods less attractive if facial deformation features and NVC labels can be considered as continuous values. One popular boosting method is Adaboost [118], which is a binary classifier known to be sensitive to outliers [220]. Adaboost was applied to emotion classification by Moore and Bowden [212] and He et al. [139]. Adaboost has been extended to temporal sequences and this is discussed in the later section on temporal model classifiers. The technical details of Adaboost are discussed in more depth in Section 4.3.1. Boosting methods often explicitly select a subset of features which is easy to interpret and this can be useful in finding the types of face deformations that are responsible for determining the label of a test sample.

A Support Vector Machine (SVM) is a kernel based learning technique that attempts to learn a decision boundary that provides the maximal separation between positive and negative samples (Vapnick [318]). In the input feature space, the boundary is non-linear and allows many complex problems to be addressed but at the risk of over-fitting, resulting in lower performance. The boundary model is based on adding weighted kernels centred at particular training samples. A test sample's distance from the boundary provides a mapping to a SVM space in which the problem is linearly separable. SVM was originally formulated for binary classification and have since been extended in various ways, including for regression, which is used later in Chapter 7. The SVM method is popular and has been applied to facial analysis many times. It has been shown to be effective in emotion classification [125, 72, 284, 213, 308, 303, 152, 339], classifying pain or no pain [18] and on AU based expression recognition [300, 123]. However, SVMs do not provide an intuitive way to interpret their internal model, or to easily determine which features are most relevant. The technical details of the SVM is discussed in more depth in Section 4.3.2. A Relevance Vector Machine (RVM) is a kernel technique that is closely related to SVM but uses Expectation Maximization (EM) to find a model to provide probabilistic label prediction. However EM does not necessarily find a globally optimal model but still is an effective classifier that results in a sparser model and is often seen to result in a higher performance when compared to SVM [224, 336]. RVM was applied to classifying brow action samples as posed or spontaneous by Valstar et al. [317]. Support Vector Data Description (SVDD) is another kernel method but in

this case addresses the problem of modelling a single class's distribution. This was used by Zeng et al. [343] to model samples in one class labelled as "emotional" and this was used to classify test samples as either emotional or unemotional. This one class approach is useful if it avoids the problem of attempting to model a class which has a complex feature space occupancy.

A decision tree classifier learns a graph of simple rules that recursively divide the feature space along axis parallel planes [40]. Decision trees are fast to train and apply, they provide a simple way for human inspection of the internal model and allow feature relevance to be determined. Decision trees were applied by Hillard and Ostendorf [140] to agreement and disagreement classification in audio. Decision trees have been extended to random forests [41] in an attempt to improve performance. Random forests use an ensemble of decision trees with each tree trained on a different subset of features, using the concept of bagging. Random forests produce a verbose model that is difficult to interpret, in comparison to decision trees, but relevant features can still be evaluated. Random forests have been applied to emotion classification on posed data [107].

While some approaches to facial analysis have used human designed feature extraction , it is uncommon to attempt to manually create a classification model. However manual classification rules were created by Pantic and Patras [234] to classify AU based facial expression. This allows human technical and intuitive abilities to be applied to the task to produce a tailored model, but this might not be appropriate for complex feature space distributions or in cases where the relevant features are unknown.

The classifier methods in this section have been concerned with samples having a single feature vector. The following section considers classifiers that specialise in modelling temporal and sequential problems.

2.5.2 Temporal, Model Level Classifiers

Temporal model classifiers are trained on samples that have been encoded as features that are an set of vectors. An ordered sequence of vectors is often based on sequential audio features, video frames or gestures in a clip of limited duration. This discussion will focus primarily on video classification, which is the topic of this thesis. Sequential

temporal model classifiers attempt to model the temporal variation of features for each class. This model can then be used to predict a label for test samples. The most popular sequential classifier is the Hidden Markov Model (HMM), which assumes a process can be characterised by transitions in a hidden state model. The transitions between hidden states are assumed to have a Markov property, which means the transition to the next state depends on the current hidden state but does not depend on previous hidden states. Each hidden state is associated with an emission model, which maps a hidden state onto a distribution of observable states. A separate HMM is trained for each class, including a class specific hidden state transition model and an emission model. For gesture recognition, the hidden states are usually discrete labels, and the emission model is typically a Gaussian distribution. If features that encode facial expressions are used, each hidden state corresponds with a distribution of facial expressions. A particular class would be characterised by the transitions from one distribution of facial expressions to the next.

HMM is an elegant theory which has been successfully used in many applications. However, there are a number of disadvantages arising from model used and assumptions that are made. The hidden state model usually allows for self transitions to remain in a state. Because of the Markov property, the probability of remaining in one state reduces exponentially in time, and it is difficult to say if this would be appropriate for NVC, or not. The emission model is assumed to be a Gaussian distribution which, for facial features, makes certain assumptions as to the properties of facial expression. Many of these issues can be addressed by increasing the number of hidden states in the model, but this quickly increases the number of unknown parameters in the transition model and the emission model. HMMs typically operate best when there is a large quantity of training data available. The number of hidden states and the topology of the transitions are difficult to determine without manual adjustment, which makes the application of HMMs rather time consuming and heuristic. Also, HMMs use discrete class specific models and these cannot directly be used to predict continuous value labels.

HMMs have been successful in the other form of human communication: verbal language recognition. Speech recognition often considers only the audible component

but some studies have combined audio and visual facial analysis with HMMs to improve recognition [252, 221]. Language recognition can even be attempted with visual only features, although previous studies have been restricted to a limited grammar [190, 129, 205, 271]. Another quasi-verbal mode of human communication is sign language, to which a HMM classifier has also been applied [294]. Although almost entirely visual, sign language is generally not regarded as NVC because multiple gestures are combined to form complex meanings. The difference between sign language and NVC is that sign language has a grammar while NVC does not. Emotion is also thought not to have a grammatical structure [57]. This lack of grammar makes it unlikely that HMMs will be particularly effective for NVC recognition. On more constrained problems, such as posed emotion or posed facial expression that begins and ends on neutral expression, a clear pattern of expression onset (appearance), peak and offset (disappearance) [91] can be seen. This may serve as a grammar of sorts, which can be used by a HMM in facial expression classification [179] or emotion classification [57, 218]. However, naturalistic emotion and NVC does not necessarily begin and end with neutral expression and this reduces the consistency on which sequential temporal model classifiers depend. el Kaliouby and Robinson [100] use Dynamic Bayesian Network (DBN)s, which are a generalisation of HMMs to classifying mental states. Their approach integrates modelling of different temporal and spatial scales into a single classifier.

Adaboost was extended to consider temporal sequences as the TemporalBoost classifier and was tested on facial gestures by Smith et al. [291]. This technique considers frames in a window of variable length, ending with the last frame of the sample. For each boosting iteration, the binary input features in the variable window have a logical "AND" and "OR" operator calculated, which are the values used for boosting. The variable window size is optimised to minimise the prediction error by gradient descent. As with most boosting methods, the trained model provides informative feature relevance information. Although this classifier considers multiple frames, it does not model how features vary in time in any direct sense. In testing, the order of frames in the window is not considered. Also the classifier is limited in that consecutive frames in a window must be considered, and the window end must correspond to the end of the clip, which makes it inappropriate if significant events occur at the start of a clip. As

with Adaboost, TemporalBoost is limited to binary inputs and binary labels which makes it less relevant to continuous value problems.

A Conditional Random Field (CRF) is a class of machine learning tools that, unlike conventional classification, which considers test samples independently, the label assignment of each sample considers the labels of "nearby" samples, which when applied to temporal data corresponds to nearby in time. In classification of video, this can enable a video frame to be classified based on adjacent frame labels. Hidden CRF extends this with the addition of a latent, unobserved state. This method, like a HMM, has discrete state transitions that are usually considered over short time periods. This may not be suitable for gradually varying human emotion and NVC signals. Morency used CRF to predict and recognise the behaviour of a human listener [215]. Bousmalis et al. [38] used Hidden Conditional Random Field (HCRF) to classify agreement and disagreement in political debates. This was later extended to Infinite Hidden Conditional Random Field (iHCRF) [39], which is capable of automatically finding the optimal number of hidden states. Rudovic et al. [268] extended the CRF-like Conditional Ordinal Random Field (CORF) method [162] to take account of the ordinal relations between onset, apex and offset for AU recognition.

Sequential classifiers consider samples that contain a set of ordered video frames. However, the following section will consider classifiers that specialise in classification of unordered sets, which may be appropriate for "grammar free" NVC.

2.5.3 Multiple Instance and Trajectory Classifiers

Multiple Instance Learning (MIL) considers classification of a set or "bag" containing one or more vectors or "instances" [196]. This may be applicable to NVC recognition, because it is possible that the presence of a single face configuration to be important in determination of the predicted label. If a particular area of feature space is critical to determine the label, it is known as a "concept", which corresponds to a specific face configuration. However, MIL places minimal constraints or assumptions on the nature of the data, so there is no common MIL approach. There are rather many different MIL algorithms that have been proposed for domain specific problems. Also, it is unclear

whether a single exemplar or concept is sufficient to encompass a range of naturally occurring NVC signals. Tax et al. [304] applied MIL to emotion to find concepts that correspond to each class and applied the concept model to emotion classification. The face configurations corresponding to these concepts were not shown in the paper.

Time varying facial expression can be represented as a trajectory in feature space. Instead of attempting to model a class's trajectories' temporal pattern, Akakın and Sankur [10] directly compared trajectories using a modified nearest neighbour approach. The mean and median distance from a test trajectory to a class's trajectories was summed to produce a similarity score. As with conventional nearest neighbour classification, this method can be sensitive to irrelevant features, so this approach may benefit from feature selection.

The discussion so far has focused on papers that apply a classification method to facial analysis. The next section discusses papers that compare different classification methods for a specific application.

2.5.4 Published Comparisons of Classifier Methods

Various classification methods have been discussed, as well as their properties with respect to facial analysis and NVC. However, it is difficult to discuss their relative performance without studies that directly compare classifiers on the same task. The best classifier for a problem is task specific, so these comparisons on emotion recognition provide circumstantial evidence that certain classifiers will be effective for NVC recognition. However, this expectation is only by analogy and needs to be confirmed experimentally. Emotion and NVC recognition are often interrelated but they are distinct problems.

Sun et al. [298] compared many classifiers for emotion recognition based on shape based, model fitting feature extraction. Their comparison included various Bayesian classifiers, variants of decision trees, k-Nearest Neighbour (kNN), SVM, etc. The authors were surprised that k-nearest neighbour classifier was most effective for this task. This suggests that the decision boundary between emotion classes is non-linear. Bagging and boosting, as an additional measure, was found to be beneficial to statistical model based

classifiers. The large quantity of training data available for this application also makes kNN more suitable in this case. Shan et al. [285] used various classifiers for emotion recognition based on appearance based LBP and Gabor features. The classifiers used were nearest neighbour, SVM and Adaboost. In this case, SVM was the most effective method, with nearest neighbour being the worse in performance. As well as providing another example of how different classifiers are effective for different tasks, the authors stress that effective feature extraction is critical to achieving good performance. Various SVM kernels were compared and the Radial Basis Function (RBF) kernel was found to have the best performance and generalisation. Wöllmer et al. [333] used audio features to classify emotional activation and valence. CRF performance exceeded SVM performance for classification. However, a regression approach to this task may be more appropriate than the use of classification.

Other comparisons of classifiers for facial analysis have been conducted that do not use emotion labels. Petridis et al. [244] compares static and temporal model classifiers for smile classification based on shape features. When considering frame based features, there was no significant difference between sequential and non-sequential classifiers. For window based features, static classification exceeds the performance of sequential classification. This suggests that feature extraction that encodes temporal information is more important than temporal modelling by a classifier for this task. Escalera et al. [106] used appearance and detection based features on conversation dominance labelled data. The compared classifiers were Discrete Adaboost and SVM (with RBF and linear kernels) with Adaboost having the best performance and generalisation. Tax et al. [304] used tracking based features to classify facial expression. Several classifiers were used including Linear Discriminant Analysis (LDA), diverse density [196] (a MIL method), a MIL clustering based method, HMM and CRF. The performance for several different AUs is shown but there is no single classifier that is optimal for all. In some cases a sequential classifier is useful, sometimes a MIL classifier is optimal and in other cases, a non-sequential classifier is best. A similar mixed result of performances was found by Akakın and Sankur [10] in use of various features to recognise head gestures. Their results are discussed in more depth in Section 4.9, because their paper is partly a response to results in Chapter 4. There was no clear winner between sequential and MIL-like

approaches. However, they found that better performance could be gained by decision level fusion of different classifiers. Pfister et al. [249, 248] used a temporal appearance based feature to attempt to classify masked emotion from microexpressions. In this study, Multiple Kernel Learning (MKL) provided a performance advantage over SVM in classifying emotion into present or not-present. However, random forests were best at detecting the presence of micro-expressions from a background of neutral expression. The experimental conditions of this study were different from natural social situations, so it is hard to draw firm conclusions on the best approach to NVC recognition. Rose [263] compared multi-class classification to single class recognisers for expression recognition and found that linear discrimination and kernel based multiclass classifiers was most effective. Rudovic et al. [267] compared RVMs, SVMs, linear regression and Gaussian Process Regression on the CMU Multi-PIE facial expression database and found that RVMs combined with pose normalisation was most effective. Buciu et al. [44] compared the cosine similarity matrix with SVMs for expression recognition, based on various ICA approaches. They found that different ICA methods reduced mutual energy in basis images, which was negatively correlated with recognition performance.

2.6 Conclusion

There are a few conclusions that can drawn from the literature. There is no single approach, be it a feature extraction method, temporal analysis method or classifier that is suitable for all problems. There are certainly differences in popularity of classifiers, with HMM and SVM being dominant in facial analysis, while decision tree based methods are relatively rare. It is also frequently observed that for a set of behaviours, a subset is easier to automatically recognise than others. The next chapter describes the collection of a corpus of informal conversations that is suitable for training an automatic system.

Verbal and nonverbal activity is a unified whole, and theory and methodology should be organized or created to treat it as such.

Kenneth L. Pike

3

The LILiR TwoTalk Corpus and Annotation Data

NVC occurs as a component of almost all forms of human communication. In order to study it, it is usually convenient to record a representative sample of human communication for later analysis. This set of data is called a "corpus". The observations in the corpus are usually labelled by a group of observers or annotators. The manner of recording and the type of annotation is dependent on the behaviour under investigation. This chapter describes the collection of a new corpus that occurs during informal conversations. Minimal experimental constraints are used in order to retain the natural and spontaneous characteristics of informal conversation.

The corpus described in this chapter has been named the LILiR TwoTalk corpus[1]. At the time this work was conducted, there were limited appropriate data sets that were

[1]The name derives from this work being associated with the EPSRC LILiR project.

publicly available. Corpuses have been recorded in various situations with a range of spontaneity and naturalness. The most viable candidate was the AMI meeting corpus [50]. However, only a portion of this corpus is naturalistic and it was not suitable for the feature tracking method employed in Chapter 4 and subsequent chapters. This chapter describes a new corpus that was designed specifically to fit the requirements of this study.

Posed data differs from spontaneous data in many ways because NVC signals are dependent on social context[2]. Informal conversations can be recognised in other cultures because they have some characteristics that are cross-cultural. The choice of a common, reproducible social situation is attractive for cross cultural study. The social situation of informal conversation is easy to organise and reproduce experimentally. However, spontaneous data is challenging to annotate because of the NVC signals being sparsely distributed throughout lengthy videos. Also, there is no clear application for informal conversation behaviour recognition at this time apart from further improving our understanding of human behaviour. Previous annotation approaches have focused on encoding a subject's internal state, emotions, gestures, dialogue acts, social relationships, topic, attention or expressions. The LILiR TwoTalk corpus uses annotation labels that encode the communicative non-verbal behaviour [93], including both the verbal and NVC aspects.

The main contributions of this chapter are:

- a new corpus of informal conversations between pairs of people which is suitable for the study of NVC signals,
- an annotation set of communicative non-verbal behaviours. The annotation was performed using a new set of NVC quantised, dimensional labels,
- inter-annotator agreement of the collected data was analysed and
- co-occurrence of NVC signals was found and

[2]Even if posed data is considered as having no social context, the absence of a social context is a factor in NVC expression.

- the recordings and annotation data are publicly available[3].

The next section provides an overview of recording conditions, annotation systems and related work. The recording of the new corpus is described in Section 3.2. Section 3.3 describes the questionnaire used by the annotators. Section 3.4 describes how multiple annotators rated the corpus video samples. Demographics of the annotators are described in Section 3.5 and Section 3.6 investigates patterns occurring in the annotation data.

3.1 Related Research

This section describes the creation and use of corpuses that can be used as the basis for computer based analysis. The most significant social situations and annotation systems are discussed, as well as the existing data sets.

3.1.1 Social Context

Acted corpuses are convenient to use because the samples have a predetermined ground truth and little recording time is wasted on uninteresting behaviour. It is impossible to predict or control the specific behaviours that will occur in naturalistic behaviour, and therefore videos require annotation to determine the labels. The sections of the video that are of interest to researchers may be unevenly distributed. Contrived situations, such as role play, tasks and games are an intermediate approach, in which participants are guided by the experimenter to maximise the useful content and allow for natural reactions to unnatural stimuli. As discussed in Section 2.1, the situation in which a corpus is recorded affects the behaviours that occur.

Historically, a large amount of emotion recognition research has been conducted on acted data sets, in which participants are told to express or pose particular behaviours, or the behaviour is expressed in a contrived or rare social situation. Often, a sequence starts and ends with a neutral expression. However, spontaneous emotion can change

[3] http://www.ee.surrey.ac.uk/Projects/LILiR/twotalk_corpus/

without transitioning through a neutral expression. Novel methods continue to be proposed for acted data sets [55], including: BU-3DFE [213], JAFFE [139], Mind Reading DVD [100] [292] and GEMEP-FERA [314]. Emotion recognition based on basic emotions is generally considered a solved problem [316]. Many studies are based on elicited emotional responses from subjects that are interacting with a device being controlled by the experimenter [7], e.g. viewing videos [298, 249], interacting with computer characters (e.g. SAL) [334] or a robot [284]. This method is also referred to as a "Wizard of Oz" situation. Further naturalism is added by having a social situation with two or more humans participating in a task. Data can be recorded in a game environment, such as EmoTaboo [342], interviews [69] or in a niche social situation, such as speed dating [193]. Contrived social situations, in which participants spontaneously react to an experimenter designed social situation, include staged interviews [343] or meetings, such as in the majority of the AMI meeting corpus [50] for which approximately "two-thirds of the data has been elicited using a [role-play] scenario" [1]. Few studies consider social situations that are not contrived or goal based activities. Almost all studies occur in the laboratory, due to the practical difficulty of recording natural social situations. Controlled situations can be useful for data collection if the automatic system is intended to be deployed in such an environment.

Informal conversations are used throughout this thesis. An informal conversation is a common social situation and one which almost everyone experiences on a daily basis. This context is also referred to as "casual conversation" or as "chatting". These conversations are usually relaxed, unfocused discussions about trivial matters. Eggins and Slade [89] defines casual conversation as "talk which is NOT motivated by any clear pragmatic purpose". This social context has not received much attention from linguists or from the human behaviour recognition community. Humphrey claimed that casual chatting can be recognised across cultures because of the activity's characteristics [145] which are:

- being informal,
- lacking focus,
- containing haphazard reiteration and

- having "topics of conversation crumble away in the compulsion of people saying what they can't help saying".

Automatic recognition of NVC in informal conversations is attractive for a number of reasons. Informal conversation is a specific social situation that is relatively easy to replicate (specifically, the social context can be staged relatively easily). It is also commonly occurring, cross cultural and occurs in almost all social groups.

However, there are potential drawbacks compared to other approaches:

- the frequency of strong emotion and intense NVC is relatively low,
- there are times in which the participants are passive, which contains little information of interest and
- labelling must be performed by annotators.

The annotation of the data is discussed in the next section.

3.1.2 NVC Annotation, Questionnaire Design and Labels

Annotation uses human observers to review and provide judgements regarding the content of the corpus. Video clips are viewed by each annotator and rated based on questions set by the experimenter. The purpose of annotation is to record the way the corpus is perceived by the annotators and thereby provide a basis to study the content of the corpus.

Many factors influence the perception of NVC signals (see Section 2.2). For annotation of a corpus, these factors are still present but can be somewhat controlled. Studies have focused on the annotation perception issue, in an attempt to reduce inter-annotator disagreement and to improve the quality of the data. Reidsma claimed that inter-annotator agreement is caused by poorly chosen annotation concepts and annotation schema, clerical errors, lack of annotator training, as well as context [258]. His thesis is currently the broadest review of the annotator agreement issue. Experts can be more

consistent than untrained observers for some annotation tasks, e.g. high quality FACS [83]. Annotation can be improved by showing the video leading up to an emotion [101], as well as showing them the entire corpus before starting to annotate [142]. Studies have noted that inter-annotator agreement was lower for stylised emotion than for spontaneous emotion [32, 8]. Annotation labels that require less interpretation might be thought of as advantageous because they have higher inter-annotator agreement [111], but this avoids the problem of perceptual differences which needs to be addressed for effective NVC recognition.

Annotation of emotion data sets have often been performed by multiple observers. The annotators use a task specific encoding system that is selected or designed by the experimenter. These annotations are usually combined to form a consensus score, either by taking the majority vote in the case of discrete classes [284, 106], or taking the mean in the case of dimensional variables [333, 218]. In this case dimensional is defined as "the range over which or the degree to which something extends" [2]. This is done to reduce the effect of different interpretation among the annotators and emphasise the generally agreed content of the corpus. However, this attempt to minimise the role of interpretation differences makes the ground truth differ from the individual human observations. A less common approach is to consider subsets of annotators and model them individually. A subset of annotators that had inter-agreement was modelled by Reidsma and op den Akker [260]. Groups of annotators can be collected and handled separately, as in the case of naive and expert annotators [83]. Although judgement based annotation is almost universally used, a few studies have used self assessment [193] or a combination of self-assessment and annotator judgements [142].

There are many pre-existing annotation systems that encode facial expression, emotion, mental states, affective state, gesture, dialogue acts, social relationships, attention and communication. These systems can be broadly grouped into four classes: those that assess the internal state of a person, a person's physical behaviour, social dynamics between people and those that describe the meaning of specific actions. Emotion labelling is one of the most popular facial or mental state labelling systems. The most common emotion labelling system is based on discrete classes, occasionally using the original Ekman 6 basic emotions [57]. There are no commonly agreed set of emotions

and the choice of appropriate labels would depend on the intended application. Discrete emotional classes cannot comprehensively cover all emotional states and instead focus on episodic occurrences [66] while ignoring pervasive emotion. Pervasive emotions are emotional states that are routinely experienced in life but not present in an emotionless state.

Emotion labelling has often been expressed in a dimensional, abstract 2D space such as activation and evaluation [280] or valence and activation [65, 182, 208]. In this case, abstract means "non-prototypical" and "defined in terms of natural language" [160]. It is unclear how many dimensions are necessary to faithfully encode human perception of NVC or emotion. A 2D space may not be enough to encode all emotions unambiguously [116]. Schröder et al. claimed that emotion could be effectively encoded using only 2 dimensions [280] in a system such as "FeelTrace", but they admit there was ambiguity in distinguishing between anger and fear. However, dimensional encoding is not limited to these labels: Ashraf et al. [18] used a dimensional scale for rating pain, Mikels et al. [210] collected emotional annotation data for images using *fear*, *sadness*, *disgust*, and *anger*, each measured on independent, 7 point Likert scales and Ball and Breeze [23] using dominance and friendliness dimensions to encode personality. Dimensional scales have also been used to rate toddler behaviour [186], facial action intensities [110], happiness and sadness [144], classroom interactions[4] and non-verbal behaviour [113]. The diversity of labels used in dimensional systems is broad to cover the variety of scientific problems to be addressed.

The AMI corpus includes many types of annotation labels including dialogue acts, attention and specific gestures [50]. The annotation was later expanded with dominance [13] and emotion. Devillers [77] used a different labelling system based on appraisal theory [277], in which the emotion stimuli are rated rather than the mental state.

Various encoding systems have been discussed but the choice of the most appropriate system is dependent on the type of behaviour of interest to the experimenter. Annotation labels can focus on either the internal mental state or the intentional communication but not usually both. An example of labels that focus on internal states is

[4] http://www.teachstone.org/about-the-class/

the Mind Reading corpus, originally created to assist autistic observers to recognise a mental state in others. This was applied to automatic recognition by el Kaliouby and Robinson [99] for a subset of labels: agreeing, concentrating, disagreeing, interested, thinking and unsure, with the emphasis on mental state. Afzal et al. used similar affect labels for recognition [8]. Pain has also been used for annotated and automatic recognition [18, 189].

Annotation labels which focus on verbal communication meaning can also be applied to NVC. Hillard and Ostendorf [140] performed classification using agreement and disagreement labels on intentional verbal utterances. Zara et al. [342] used high level groupings of verbal communication acts in EmoTaboo, some of which correspond to a meaningful communication. Bavelas and Chovil [28] counted the frequency of meaningful facial gestures.

Another group of human behaviours that are of interest to researchers is facial expression. Facial expressions are an externally observable movement of the body which require less interpretation than emotion to annotate. This is in contrast to emotions which largely occur in the mind and only sometimes manifest themselves in behaviour or expression. The most popular method for encoding facial expression is the Facial Action Coding System (FACS) [95], in which facial Action Unit (AU)s correspond to sets of muscles. FACS is widely used for labelling expressions and has been used as the basis for regression systems ([276]), although most papers only use a subset of FACS. However, FACS annotation exhaustively encodes expressions, requires trained observers and is very time consuming to perform. The encoding is typically based on binary classes and does not consider the intensity of expression. Others have used expression labels that did not use the FACS system, for example Kanaujia et al. [157] labelled nodding and blinking.

There are other facial analysis labelling approaches that do not fit with the previously discussed groups but have some relation to NVC. Deception and truthfulness modifies the perception of communication and have been used as labels and in recognition [309, 249]. The outcomes of speed dating was labelled and recognised [192] and this again is largely based on perception of NVC, but the labels themselves do not correspond

to meaningful communication. Given that the situation in which data is recorded is significant, the next section discusses why a new naturalistic data set is needed for studying occurrences of meaningful NVC signals.

3.1.3 The Need for a Naturalistic Corpus

There has long been criticism of the study of social phenomena in a laboratory environment. Although a laboratory is intended to assist the control of experimental variables, Argyle [16] claimed that too often study subjects "sit in cubicles by themselves, watch flashing lights and press buttons; often there is no verbal communication, no NVC, no real motivation, and there are no situational rules". Moving a social situation from its normal location to the laboratory can have significant effects, due to the subject's knowledge that they are being recorded [30] but ethical and practical considerations make this fact hard to conceal [117]. Posed and spontaneous data also have significant differences. If meaning in communication is the subject of study, context is significant which makes posed or over simplified data unsuitable [28]. However, not all human related research needs to be naturalistic and researchers need to assess the suitability of any database [85]. Many emotion recognition systems have used posed data and this seems unlikely to change. However, different recording conditions have been shown to result in different behaviours. If an automatic system is trained based on a corpus and deployed in a different environment, it is quite possible that human behaviours will be significantly different and the automatic system will have poor performance. For example, the timings of natural and posed emotions are different [58]. This difference is so significant that posed and spontaneous examples of emotions can be manually and automatically distinguished [317, 248]. Strong emotion is rarely expressed or is expressed in unusual circumstances, making recording of naturalistic examples difficult [67]. For natural data, the information content is unevenly distributed in time [69]. Rapid emotional transitions are more common in natural data than in posed data [208, 28]. Rapidly changing emotion has higher annotation demands and less predictability in recognition. Cowie [67] reviewed these issues in the context of creating databases suitable for human behaviour and concluded the challenges for recording and annotating such a database are significant, but stressed that these issues must be

addressed to make progress. An alternative to using naturalistic data as the basis for automatic systems is to use data "of a kind that might have to be dealt with in an application" [68]. In recent research, these types of datasets are becoming more popular. It is likely that some applications of NVC recognition require the use of naturalistic data and this remains the focus of this thesis.

3.1.4 Existing Data Sets

The majority of facial analysis and human behaviour data sets have focused on emotion or expression recognition. As discussed in the previous section, there is a need to use natural data, therefore many of the existing data sets are not appropriate for NVC recognition. The existing naturalistic, public databases will be reviewed and the reasons for recording a new data set will be discussed in this section.

- *Belfast Naturalistic Emotional Database* is a corpus using both television programmes and interviews [66]. The clips are between 10 and 60 seconds in duration. The variability of social context would make expression of NVC rather diverse. Also, some of the social situations are rare, such as being an interviewee on a television programme.

- *EmoTV* comprises of 89 television interview monologues. The duration of video clips varies between 4 and 43 seconds. Again, the variability of social context is an issue, as well as the low quality of analogue broadcast TV. [77]

- *FreeTalk* is a four person conversation which was not limited in topic. Participants remained seated throughout. The conversation was conducted in a laboratory. This corpus is similar to the one presented in this chapter but FreeTalk was publicised after the TwoTalk corpus was recorded and used [48].

- *D64 Multimodal Conversational Corpus* used unrestricted, multi-person conversation in a domestic environment [227]. This corpus is a significant improvement on previous data sets in that it recorded casual conversation outside of the laboratory environment. The subjects could move around or leave as desired. This

data set was also publicised after the corpus in this chapter was recorded and used.

Although not the primary focus of this thesis, existing non-naturalistic datasets include:

- *Canal9* is a series of broadcast television political debates between 2 or more participants and a moderator. This corpus is discussed in more detail in Section 7.7.
- The majority of the *AMI Meeting corpus* is a series of recordings of role play meetings in a group of 4 people. Approximately two thirds of the meetings are role play scenarios, and the remainder are naturalistic. People occasionally moving from seated to standing (Figure 3.1).
- *EmoTaboo* is a set of recordings of role play games between two participants.
- *MMI* is a searchable database of posed and elicited emotions.
- *Mind Reading* corpus is a library of short, silent videos of mental states performed by actors. This was originally produced as training material management of autism spectrum disorders. This corpus is described in more detail in Section 7.8.
- *Sensitive Artificial Listener (SAL)* is a corpus of elicited emotion based on a human interacting with a computer character that exhibits one of a set of personality types.
- *Green Persuasive Dataset* is a series of recordings of a role play situation between an experimenter and a volunteer with the discussion topic focused on environmental issues. The experimenter attempts to persuade the volunteer to adopt a more environmentally sustainable lifestyle.
- *MHi-Mimicry-db[296]* is a 15 camera, 3 microphone recording of dyadic conversations. The participants were either in a political debate (34 recordings) or in a role-playing game (20 recordings).

Table 3.1: Summary of data sets used for emotion and conversation oriented research.

Corpus	Duration	Context	Participants	Labels
Belfast Naturalistic [66]	86 min labelled 12 min available	emotional interview	Dyadic	Emotion
EmoTV [77]	89 clips, 12 minutes	interview monologues	Monologue	Various
FreeTalk [48]	270 minutes	lab, conversation	4 person	Various
D64 Multimodal [227]	8 hours	domestic, conversation	4-5 people	Unknown
Canal9 [323]	42 hours	debate	2 to 4 person & moderator	Shots, ID
AMI Meeting [50]	100 hours	Role play meeting	4 people	Various
EmoTaboo [342]	8 hours	Mime game	Dyadic	Emotional Events[77]
MMI [312]	Increasing with time	Various induced	Single participant	Various
Mind Reading[a]	˜19 minutes	Posed	1 Actor	Mental state
SAL [279][b]	˜10 hours	Wizard-of-OZ	1 person with computer	Emotion
Green Persuasive Dataset[c]	videos 25-48 minutes 8 dyads	role-play	dyadic	persuasiveness
MHi-Mimicry-db[296]	˜12 hours	discussion/role play	40 people, dyadic	various

[a]http://www.jkp.com/mindreading/
[b]http://semaine-db.eu/
[c]http://green-persuasive-db.sspnet.eu/

Table 3.2: Summary of adverse factors for the suitability of existing datasets.

Corpus	Suitability
Belfast Naturalistic [66]	Staged interview, only a ˜12 min subset is available not labelled for NVC, face size is small (approximately 150 by 200 pixels)
EmoTV [77]	Not publicly available
FreeTalk [48]	Video is small and faces are approximately 18 by 18 pixels unsuitable for tracking, only publicly available recently (since 2010)
D64 Multimodal [227]	not publicly available at time of writing
Canal9 [323]	Unusual social situation, only a subset is annotated ˜10 min not continuous view of subject, low quality interlaced broadcast video, standing multiple participants may result in larger head pose changes
AMI Meeting [50]	Mostly contrived role play scenario although some meetings are naturalistic, not NVC annotated, emotion annotations are not publicly available multiple participants may result in larger head pose changes and more complex interactions, video is highly compressed, face size is small (typically 110 by 70 pixels), contains a mixture of standing and seated behaviour
EmoTaboo [342]	Not publicly available
MMI [312]	Focused on posed and induced expression not NVC Only uses a single participant and not dyadic
Mind Reading	Acted, low quality video, face only 100 by 150 pixels, focuses on mental states not NVC
SAL [279]	Human to computer character conversation rather than human to human conversation Available since Mar 2009
Green Persuasive Dataset	Contrived social situation, available since 2009, Small videos with face approximately 104 by 145 pixels
MHi-Mimicry-db[296]	staged discussions or role-playing games, labelled for facial expression not NVC, only recently available (since 2011)

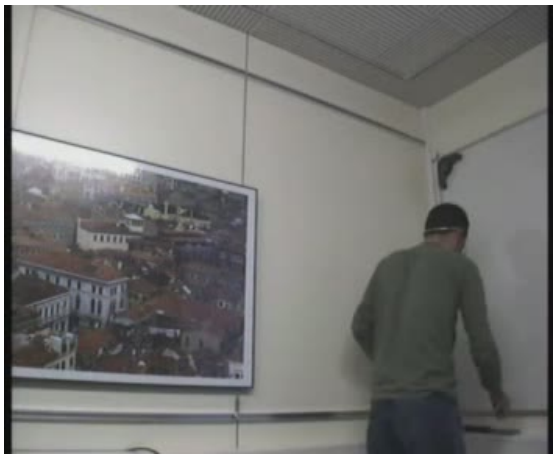

Figure 3.1: An instance in the AMI Meeting corpus of a participant getting up from a seated position and standing near a white board with their back facing the camera (in video IN1014.Closeup2.avi frame 5200). This behaviour makes facial tracking problematic.

There are many other task based or induced emotion corpuses (see Table 3.1), but they exhibit greater variability in social context, low video quality or the task being too specific make these data sets unsuitable. See Table 3.2 for the suitability of each data set. Also, many corpuses use more than two participants but limiting conversations to two persons is likely to be simpler to understand and analyse. Corpus videos in which the face has a small size can be difficult to accurately track. Corpus videos that feature more than two participants may have a larger head pose variation due to people turning to face different people during the conversation. These factors can make tracking less effective. Later chapters use Canal9 and Mind Reading for regression (see Sections 7.7 and 7.8), but this thesis is primarily focused on a new NVC corpus TwoTalk. The next section describes how the TwoTalk corpus was recorded.

3.2 Description of LILiR TwoTalk Corpus

Two participants were selected from the department and invited to a data capture session in a visual media lab. The only criteria used to select participants was the requirement to have people of roughly equal social seniority. Culture, familiarity and

Figure 3.2: An example frame from each of the eight participants. The top row, looking left to right, are participants 1008, 1011, 2008 and 2011. The bottom row are participants 3008, 3011, 6008 and 6011.

gender were not controlled in participant selection. However, these differences may affect the type and frequency of NVC signals.

The laboratory was selected as the setting to perform data capture. This choice was based on the available cameras being directly wired into a fixed, non-portable data recording system. Videos were recorded using two progressive scan, PAL digital video cameras with a frame resolution of 720 by 576 pixels and 25 Hz frame rate. The cameras were arranged to record facial behaviour, which is involved in the expression of many types of NVC [15] [215]. The face size in the video was typically around 200 by 300 pixels. The cameras were genlocked to ensure frame synchronisation. The error in synchronisation was smaller than the limits of measurement (a fraction of a microsecond). To minimise synchronisation error, cameras of the same type and synchronisation cables of the same length were used (this ensures the signal propagation from the generator to the cameras takes the same time duration). The arrangement of the lab equipment is shown in Figure 3.3. The corpus was recorded in 2009 and before the availability of affordable consumer depth cameras (which provide both an optical image and a depth map). Recent data sets have begun to utilise "two and half"D or 3D recording [108] with this type of equipment.

Once both participants arrived in the laboratory, they were given only two instructions:

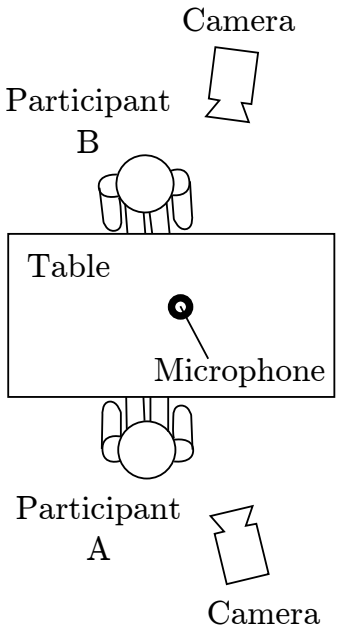

Figure 3.3: Plan view of video capture equipment arrangement.

to be seated and to communicate until told to stop. Having the participants seated reduces the amount of variation in body and head pose. Without this constraint, participants tend to turn away from the camera which makes facial tracking difficult. The experimenters were not visible to the participants during the recording. The participants were then allowed to talk without further experimenter interaction for 12 minutes. The conversations were of limited duration because the participants may begin to tire and change their behaviour. The instruction to communicate was considered necessary because the laboratory is not a normal place for socialising. The participants seemed to ignore their artificial surroundings and interacted in a natural fashion. The demographics of the participants in the four dyadic conversations are shown in Table 3.3. The number of conversations and duration was based the need to capture a range of NVC behaviours without tiring the participants. The use of eight subjects is suitable for person independent cross validation testing with the majority of the data being available for training (87.5% in training, 12.5% in test). Using the person specific LP tracker, more individuals and conversations also requires more person specific training to achieve acceptable tracking accuracy and additional resources required to organise and record the corpus. However, more participants results in an increased number of

Table 3.3: Demographics for Conversation Participants in the LILiR TwoTalk Corpus. Abbreviations: UG is an undergraduate degree. M is a master's degree. Certain entries are omitted in cases where the participant was no longer contactable.

Participant	Country of Origin		Age Years	Years UK Resident	Education	Languages Spoken	Natively
1008	Nigeria	♂	25	16	M	English	Yes
1011	British	♂	29	25	UG	English	Yes
2008	Spain	♂					
2011	British	♀	27	27	UG	English	Yes
						French	No
						British Sign	No
3008	Mexico	♂	29	5.5	PhD	Spanish	Yes
						English	No
3011	Sri Lankan	♂	25	6	M	Bengali	Yes
						English	No
						Hindi	No
6008	Indian	♂	27	1	M	English	Yes
						Hindi	Yes
6011	Ukranian	♀	27	2	M	Ukrainian	Yes
						Russian	Yes
						English	No
						French	No
						German	No

cross validation folds, which allows the standard deviation of the performance to be estimated more accurately. Less individuals in the corpus requires less training data for LP tracking (see Appendix B), but results in a smaller proportion of data available for training in cross validation. Different cultural pairings were used in each conversation, which rules out the possibility of a study of cultural differences in expression within the TwoTalk corpus. Later chapters consider cultural perception differences, rather than expression differences caused by cultural background. Example frames from the corpus are shown in Figure 3.2.

Four conversations of 12 minutes length provide 48 minutes of conversation. The con-

versation was recorded by two cameras, resulting in 96 minutes of video. For multiple annotators, the inter-annotator agreement for corpuses of emotion [258] and NVC is low. For this reason, each NVC signal clip was rated by multiple annotators to reduce the effect of person specific factors. Much of the recorded conversations contain only passive listening, with no apparent NVC displays, which is only marginally interesting. As Cowie et al. [69] observed, the distribution of signs is not uniform in human communication. Sections of recordings that contained little or now activity were manually identified and excluded. The remaining sections of video contained potentially interesting behaviours. The annotators were not informed of the particular NVC that was potentially present in clips of interest, so the NVC content of the corpus was rated entirely by the annotators. 407 clips were selected as potentially of interest. The proportions of NVCs that were thought to be present in the initial clip selection process is shown in Figure 3.4. The different NVC frequencies are due to the natural frequencies of occurrence of the various NVC signals. An additional 120 randomly selected clips were also included to increase the variety of samples in the corpus. This may include common NVC signals that were not considered in the questionnaire and also samples of null NVC expression. This increases the richness of the videos shown to the annotators and prevents them reducing the problem to a hard assignment, 4 class problem which may influence the resulting annotation data. Both the manual and random sets of clips are combined, to form a final set of $o = 527$ clips with an overall duration of 38 minutes. The clip lengths used ($L = 0.6$ to 10 seconds, average $\overline{L} = 4.2s$, standard deviation $\sigma(L) = 2.5s$) are similar in duration to Lee and Beattie's work in discourse analysis [175] (sample lengths of $\overline{L} = 4.8s$, standard deviation $\sigma(L) = 2.5s$). If only total duration is considered, this data set is somewhat smaller than other data corpuses. For example, the emotionally labelled subset of the Belfast Naturalistic Emotional Database has a duration of 86 minutes [85], AMI Meeting corpus has 100 hours [50] and Madan's speed dating corpus has 350 minutes [192]. However, these corpuses have different annotation methodologies and goals, and in many respects are not comparable with the TwoTalk corpus. Canal 9 is a large corpus (42 hours) but only a 10 minute subset is annotated for agreement and disagreement NVC.

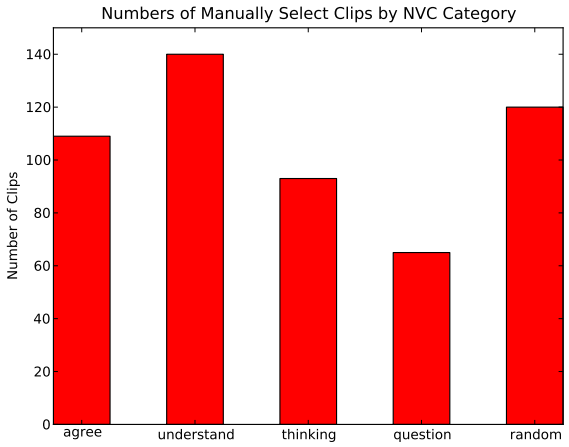

Figure 3.4: Number of manually selected and randomised clips in each of the NVC categories of interest. For *agree*, *understand*, *thinking*, *question* and *random*, the number of clips are 109, 140, 93, 65 and 120 respectively.

3.3 Design and Description of the Annotation Questionnaire

In attempting to encode NVC signals, a system was selected that encompasses as many NVC signals as possible, while making the annotation system easy for the annotators to use. In a similar manner to emotion, NVC signals change over time and, due to the flexibility of human expression and human interpretation, a vast range of communication signals are possible. However, the specific labels used for NVC encoding need to be selected as the basis for annotation.

An interpretative label is used to encode meaning in NVC. An encoding scheme can have a "categorical" or "dimensional" encoding basis [66]. Categorical systems rate events based on its similarity to a set of exemplars. Usually these exemplars are based on language that is easy to interpret (e.g. Ekman 6 basic emotions [96]). Abstract dimensional systems, which are not based on a simple similarity between an observed to an exemplar expression, attempt to represent a much broader range of events using abstract rating scales. Abstract dimensional encoding includes the commonly used scales activation and valence [76]. Given the ease of use of exemplar NVC concept labels

and the lack of any low dimensional, abstract encoding system for NVC encoding, an prototypical (exemplar) NVC basis was selected.

Multiple NVC signals may occur simultaneously and with a range of intensities. NVC encoding needs to address these possibilities. For this reason, the popular approach of using multi-class, hard assignment (mutually exclusive) labels is not suitable for capturing the nuances of NVC signals. A high dimensional prototypical approach was adopted, which considers NVC on multi-dimensional scales [171] which can independently vary. This method is relatively easy to use while accurately encoding a subset of co-occurring NVC signals.

For temporal annotation of video there are a few approaches that may be used:

- One annotation strategy is to evaluate the video using a continuous input device in real time to form a temporally continuous annotation. This thesis considers the issue of temporally continuous to be distinct from being dimensional, continuous valued labels in an NVC or emotion dimensional space, however a corpus may possess both properties as in the case of the FeelTrace system [280]. Dimensional encoding is typically used for encoding dimensional, continuous valued labels [223], although it would be possible in principle to use temporally continuous categorical encoding. Temporally continuous encoding requires specialised equipment and additional effort by the annotators.

- The alternative is to use individual clips rather than long uninterrupted videos. However, this raises the issue of how the start and end frame of clips are selected.

To control the resources expended in annotation, short clips were manually extracted from the original video recordings and used as the basis for dimensional, continuous valued annotation in this study. This study is therefore not temporally continuous.

There is no standardised, comprehensive set of NVC signals and little existing research in this area. In one of the few direct studies of NVC, Lee and Beattie [175] examined the use of gaze and Duchenne smiles in an NVC context. However these NVC signals do not have any obvious practical application. Another study, based on recognizing mental states, was conducted by el Kaliouby and Robinson [99]. The mental states

Table 3.4: Questions used in web based annotation of the LILiR TwoTalk corpus.

Question for Category	Minimum Rating	Maximum Rating
Does this person disagree or agree with what is being said? (A score of 5 is neutral or not applicable.)	Strong disagreement	Strong agreement
Is this person thinking hard?	No indication	In deep thought
Is this person asking a question?	No indication	Definitely asking question
Is this person indicating they understand what is being said to them?	No indication or N/A	Strongly indicating understanding

they studied were *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. Mental states are not necessarily expressed outwardly and can be expressed with or without communicative intent; this is in contrast to NVC which is always communicative, outwardly expressed and intentional. The criteria used in this study for selection of NVC signals to study were:

- the signal should be an intentional NVC action,
- they commonly occur in the social context used (i.e. casual chat/informal conversation),
- they may be potentially useful in one or more application and
- they may modify the meaning of the literal words used.

Based on the previous design decisions, the questionnaire is a quantised, multi-dimensional[5] encoding using Likert scales of short video clips. The questionnaire presented to the annotators is summarised in Table 3.4 and reproduced in Appendix D. In this thesis, the 4 categories are abbreviated to *agree*, *thinking*, *question* and *understand*, and refer to the content of this questionnaire. The labels are exemplar based (in a similar fashion

[5]Questionnaires can be used to collect multi-dimensional responses e.g. [20].

to the Ekman basic emotions), unlike activation and valence which are not exemplar based. This thesis uses dimensional labels that are tailored for the needs of NVC encoding. As previously discussed in Section 3.1.2, dimensional encoding is used in a wide variety of human behaviour annotation and is not limited to activation and valence.

As with el Kaliouby and Robinson, the NVC labels of *agree* and *disagree* were used because these met the above criteria. Agreement signals are among the most prevalent of NVC signals [37]. Agreement and disagreement are mutually exclusive messages, therefore these labels are expressed on a single dimensional scale. The NVC signal *thinking* was used because it is a common and distinct behaviour in the corpus, based on an informal viewing of the corpus videos. Also, the tracking method used is very effective for eye movements and gaze is known to play a role in conversation turn taking [15]. The way gaze varies during thinking is culturally dependent [206], making it a potentially interesting NVC signal when comparing different cultural perceptions of NVC (see Chapter 7). However, *thinking* is arguably a mental state, as much as a true communication signal. *Question* is included because it drastically changes the meaning of an utterance, which may have applications in multi-modal speech recognition. The inclusion of *question* was debatable, because while questions can be recognised by changes in voice intonation, it was unknown if visual recognition of a question was even possible. *Understand* is another common signal which regulates conversation and has both verbal and non-verbal components. The NVC signals of *agree* and *understand* express attitudes and are therefore social signals. While *thinking* and *question* are likely to have a role in regulating conversation flow, their exact role and significance is uncertain and their status as social signals is current uncertain.

A questionnaire can collect dimensional responses may have labels values that are either discrete (quantised) or continuous. Continuous value data may be collected by marking a point on a continuous line using either a written mark or computer input [307]. The term "continuous value" does not exclusively apply to annotations with continuous temporal traces, such as FeelTrace[6]. The concept of "continuous emotional

[6]Human behaviour may either be encoded as either discretised or continuous in terms of annotation labels: "When discretised dimensional annotation is adopted (as opposed to continuous one), researchers seem to use different intensity levels" [132]

space" is expanding beyond activation and valence, with Hupont et al. [146] using a 2D continuous value space (*evaluation* and *activation*) without considering the temporal dimension. Liscombe et al. [182] used a web-based survey to gather dimensional, continuous valued emotion annotation data for each utterance in an audio corpus. Discrete binary labels might simply be "NVC is expressed" and "NVC is not expressed". Binary discrete labels were used by Lee and Beattie [175] in their discourse analysis of NVC. Discrete labels can also be used in dimensional annotation and may be encoded using integer values. Continuous valued labels are dimensional ratings that have non-integer values and are commonly used in psychological research [223, 334, 280]. A Likert scale [180] is a common psychometric questionnaire in which a ordinal rating can be provided between two extreme choices. Taking the mean of multiple annotators results in a value that can take non-integer values and is therefore a dimensional, continuous valued variable. While some concerns have been raise as to the validity of using a Likert scale as an interval-level measure with parametric statistics [150], a number of studies have examined the use of Likert scales as an interval-level measure and found it to be a reliable research tool [256] [233]. Both Carifio and Perla [49] and Norman [226] argue that it is time to put the controversy of the use of the Likert scale as a interval-level measure in the past because of the repeated demonstration of the worth of using Likert scales with parametric statistics, such as mean and variance. A Likert scale was used in this study as this retains intermediate intensity expressions while being applicable to an Internet based survey.

3.4 Multi-observer Annotation of NVC

Inter-annotator agreement for inter-personal events is typically low. ("These annotations [of subjective corpura] often have a quite low overall level of inter-annotator agreement" [258]) This is also likely to be true for NVC perception. NVC perception is highly subjective and partly dependent on person specific, social and cultural factors (see Section 2.1). Multiple annotators are used in an attempt to reduce the effect of interpersonal variations. Other techniques to improve inter-annotator agreement were not employed (e.g training the annotators, having the annotator previewing the corpus,

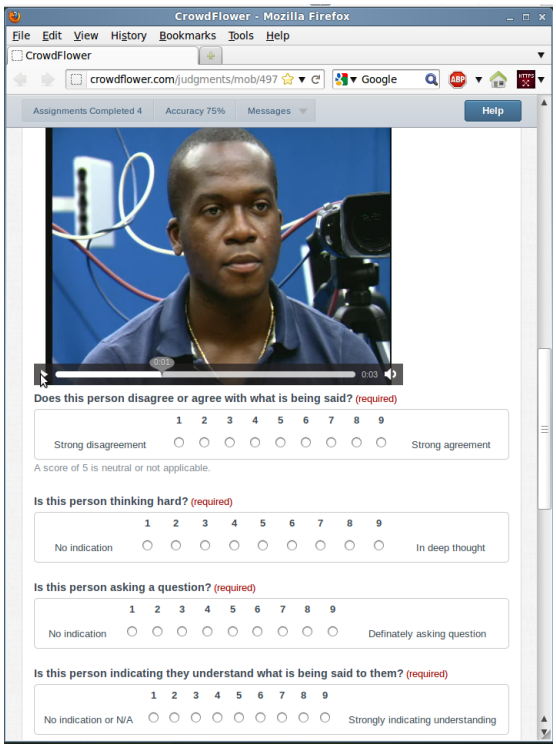

Figure 3.5: A typical page of the annotation questionnaire, which is accessed using a web browser. Each clip is annotated individually. The user has the ability to replay the video and is required to respond to the four NVC questions.

multiple ratings on individual clips by individual annotators, subject normalisation). This was due to the limitations in resources needed to conduct such a study and the technical limitations of crowd sourced annotation. The differences in perception between cultures are not considered in this chapter, but this issue is revisited in Chapter 6.

The questionnaire was presented to annotators using a computer based system. These are commonly used for annotation of video [87, 280, 302] and allow participants to complete the survey at their own convenience, rather than having to attend an organised session. A web based system was used, because the system could be remotely accessed using a web browser. A web page displayed one or more videos and each could be viewed one or more times. Under each video were the four annotation NVC categories selected in Section 3.3. The user viewed the video clip and marked their answer using the mouse or keyboard. The order of the videos was randomised to reduce the possible effect of

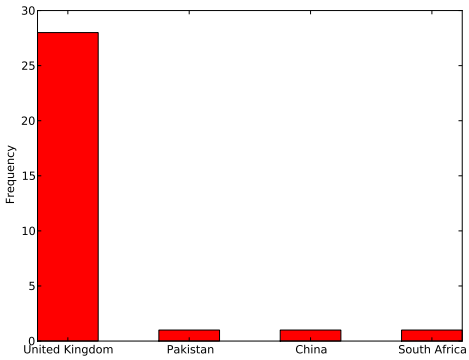

Figure 3.6: Self-reported primary culture for annotators.

the video display order on the annotation data. A typical view of the annotation web site is shown in Figure 3.5. The videos were presented to the annotators with both visual and audio information. By presenting both, it was hoped that the most natural and realistic rating would be achieved.

Allowing users to perform the annotation task with no supervision risks the participants not completing the annotation task as instructed. The annotators were unpaid volunteers and were motivated by interest, loyalty or duty and were expected to generate relatively good quality data. Another approach used is to pay people to perform the annotation but this can lead to quality issues, which are addressed in Section 6.3. Chapter 6 addresses collecting annotation data from observes based in different cultures.

3.5 Analysis of Human Annotators

Thirty one annotators participated in rating samples of NVC ($d = 31$). Because expression and perception is dependent on cultural context, gender and personality differences (see Section 2.1), it is relevant to consider the demographics of the annotators. Although all annotators were UK residents, some annotators had a separate primary cultural background. Annotators were mostly 21-30 years of age, with a science or engineering background (see Figure 3.7 and 3.8). As can be seen in Figure 3.6, the vast majority of annotators had a UK based cultural background. Annotation by distinct cultural groups is discussed in Chapter 6. The majority of respondents were male

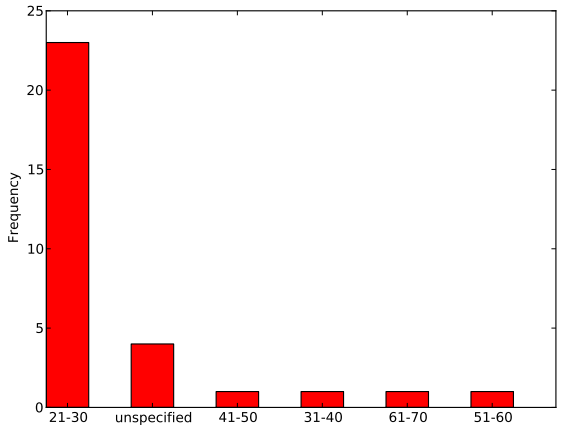

Figure 3.7: Self-reported age for annotators.

(see Figure 3.9). It might be beneficial to include a personality questionnaire in future annotation projects to enable investigation of personality in a more explicit way.

The vast majority of annotators did not rate every video clip because NVC annotation is time consuming and tedious. On average, each annotator rated 70 video clips, from a possible maximum of 527 clips. Each clip required all four NVC signals to be rated. For the NVC category *thinking*, 2182 individual ratings were provided by the annotators, distributed across the corpus's 527 video clips. This corresponds to 4.1 ratings per video clip.

Not all NVC signs occur with the same frequency or intensity. Figure 3.10 shows the distribution of NVC intensity for the annotated videos. The frequency of NVC occurrence in these videos almost certainly differs from unedited video, because the video clips in the corpus were specifically selected to include interesting NVC examples and to avoid inactive sections. *Question* NVC has a strong peak at zero, indicating the majority of video clips do not contain this signal. *Agree* and *thinking* NVC have a similar peak near "no NVC signal", but differ from *question* by having a minority of clips of intermediate intensity. *Understand* NVC has a relatively uniform frequency distribution, with some weak, intermediate and strong examples. Intense expression is rare for the NVC signals considered in this work. Combining multiple individual annotator ratings to a single vector label will now be considered.

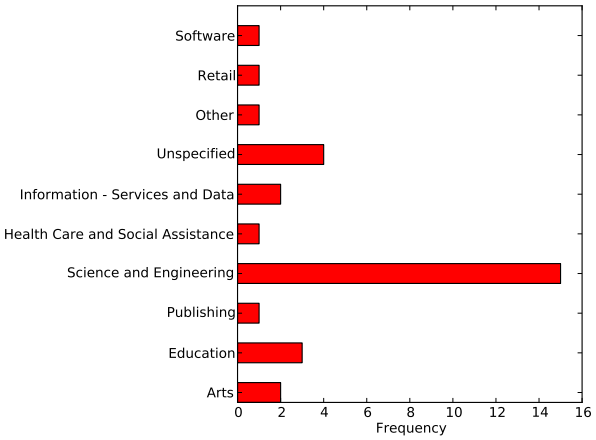

Figure 3.8: Self-reported sector of employment for annotators.

3.6 Analysis of Mean Ratings

Annotator ratings are provided for each of the NVC categories:

$$c \in N = \{agree, understand, think, question\} \tag{3.1}$$

For supervised learning, the individual ratings need to be reduced to a single consensus 4D label for each clip. This is performed by taking each NVC quantised, dimensional rating and calculating the mean (similar to [302, 334]), which results in a dimensional, continuous valued label. The vector containing the number of annotation ratings for o clips is defined as $\mathbf{z} \in \mathbb{N}^o$. The ratings for clip $m \in \{1...o\}$, annotated for NVC signal category c, is designated as matrix $\mathbf{N}_{c,m}$ of size $4 \times o$. $\mathbf{N}_{c,m}$ comprises of a set of tuples, containing ratings $\mathbf{r}_i \in \mathbb{R}$ and corresponding annotator indices $\mathbf{s}_i \in \{1...d\}$:

$$\mathbf{N}_{c,m} = \{(\mathbf{r}_i, \mathbf{s}_i)\}_{i=1}^{\mathbf{z}_m} \tag{3.2}$$

The 4 NVC categories and can be summarised into a consensus vector $\mathbf{C} \in \mathbb{R}^{4 \times o}$, $c \in N$:

$$\mathbf{C}_{c,m} = \frac{\sum_{i=1}^{\mathbf{z}_m} \{\mathbf{r}_i : (\mathbf{r}_i, \mathbf{s}_i) \in \mathbf{N}_{c,m}\}}{\mathbf{z}_m} \tag{3.3}$$

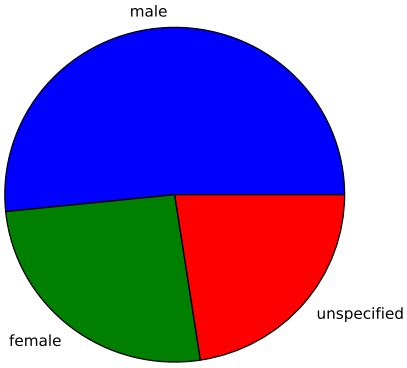

Figure 3.9: Self-reported gender for annotators.

This simplification assumes that inter-annotator differences are not significant for the intended application and that the annotator ratings are symmetrically distributed. However, human perception of NVC depends on many factors (see Section 2.1). The bulk of this thesis attempts to create and evaluate an automatic system that produces a prediction in a similar fashion to a human annotator. This chapter assumes the annotators form an approximately self-consistent group because of the demographic similarity of the annotators (see Section 3.5). Therefore, using the rating mean to determine a consensus is valid in this case. However, it is important to remember the limitations of using the mean consensus, and specifically that taking the mean of non-homogeneous groups risks over-simplification of the problem. Chapter 7 considers the case of training an automatic system when multiple distinct groups of annotators exist.

An alternative to using the mean rating is to take a set of multiple annotator rating data and use a subset of clips in which there is agreement, as done by el Kaliouby and Robinson [99]. This approach was not used, because it ignores clips in which there is low inter-annotator agreement.

The selected NVC categories do not necessarily vary independently of each other. A linear dependence of two continuously varying signals can be found by calculating the Pearson correlation coefficient ρ. The Pearson correlation coefficient ρ for vectors

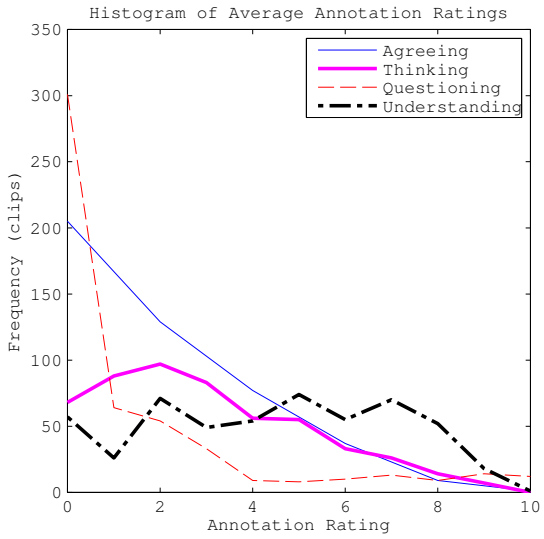

Figure 3.10: Histogram of Average Rating based on multi-annotators. Blue thin line is agreeing, magenta thick line is thinking, black dot-dashed line is understanding and red dashed line is questioning. Zero is a neutral score and 10 is strongly showing the communication signal. Disagreement ratings have been omitted.

	Agreeing	Understanding	Thinking	Questioning
Agree	1			
Understand	0.46	1		
Thinking	-0.21	-0.23	1	
Question	-0.18	-0.40	0.06	1

Table 3.5: Correlation coefficients of the mean ratings $\mathbf{C}$ for each category.

$\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, is defined as (adapted from [329], for a population, $i = \{1..n\}$):

$$\nu(\mathbf{x},\mathbf{x}) = \sum_{i=0}^{n} {\mathbf{x}_i}^2 - n\bar{\mathbf{x}}^2 \tag{3.4}$$

$$\nu(\mathbf{y},\mathbf{y}) = \sum_{i=0}^{n} {\mathbf{y}_i}^2 - n\bar{\mathbf{y}}^2 \tag{3.5}$$

$$\nu(\mathbf{x},\mathbf{y}) = \sum_{i=0}^{n} \mathbf{x}_i\mathbf{y}_i - n\bar{\mathbf{x}}\bar{\mathbf{y}} \tag{3.6}$$

$$\rho(\mathbf{x},\mathbf{y})^2 = \frac{\nu(\mathbf{x},\mathbf{y})^2}{\nu(\mathbf{x},\mathbf{x})\nu(\mathbf{y},\mathbf{y})} \tag{3.7}$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean values of $\mathbf{x}$ and $\mathbf{y}$ respectively. The value of ρ gives an indication of the correlation between two signals. $\rho = 1$ is perfect positive correlation, $\rho = 0$ indicates no correlation and $\rho = -1$ indicates perfect negative correlation. The correlation between different NVC signals is shown in Table 3.5. The highest magnitude correlation score is between *agree* and *understand*, with a score of $\rho = 0.46$. This is a relatively weak correlation but significant enough to say there is a relationship between these signs. This also confirms our intuitive expectation, because if a person wishes to indicate agreement, this necessarily implies they also want to convey that they understand. The next highest magnitude score is for *understand* and *question* at $\rho = -.40$. Being a negative correlation, this reflects that these signals are partially mutually exclusive; when a person is asking a question, they are unlikely to be conveying that they understand (and visa versa). Finally, the lowest magnitude score is for *thinking* and *question*, having a correlation of $\rho = 0.06$. This implies these NVC signals occur relatively independently of each other.

This analysis is interesting, because if there are many NVC signals that are interrelated, it can imply that dimensionality reducing techniques can be applied without loss of information. The NVC signals *agree* and *understand* often co-occur, therefore an approach using a single discrete class label would be problematic for NVC recognition.

The ratings for each annotator are compared to the mean of the other annotators. This provides a measurement of inter-annotator agreement. Figure 3.11 shows a histogram of the number of annotators at different levels of agreement with the cultural consensus. This shows that some annotators are in close agreement with the cultural consensus,

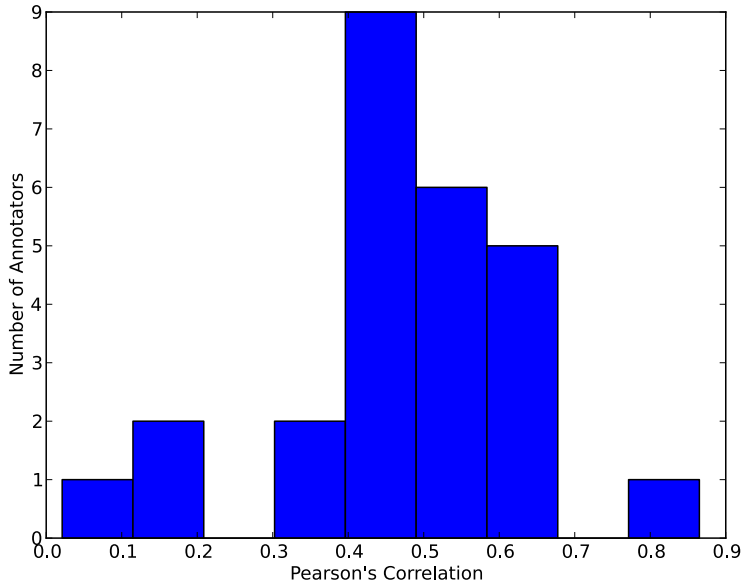

Figure 3.11: A histogram of Pearson's correlation between an annotator's ratings and the mean of all other annotators' ratings. Some annotators did not rate enough data to compute a valid correlation and were excluded from this figure.

while others are less so. For all 31 annotators, the average correlation with consensus is 0.47. Inter-annotator agreement is discussed again in Section 6.3.1 in context of removing outlier annotators.

Correlation is used as the primary measurement of agreement of survey data in this study. The use of the popular Cronbach's Alpha is not a suitable measure of agreement or internal consistency [289, 261, 290], with Green and Yang calling for it's general use should be discouraged [130]. This metric is therefore not used in this thesis. Also, the habitual use of a particular threshold (such as the commonly used value of 0.7 or 0.8) to determine if a questionnaire validity risks over-emphasising the problem of random noise, to which machine learning techniques are somewhat robust, and ignoring the problem of systematic errors by the annotators (See Chapter 3 of Reidsma [257]). Reidsma argued that using corpus data with a higher inter-annotator agreement generally leads to better performance in automatic recognition.

3.7 Conclusion

This section has described a corpus that comprises of recordings of two person, informal conversations. Minimal experimental constraints were used to maximised the natural and spontaneous character of the social situation. Annotation was performed on video clips to encode the conversation meaning. The annotators were shown the video clips with audio. The data was collected based on quantised, dimensional Likert questions. Based on this, a consensus label based of the mean of multiple annotators was calculated for each question which resulted in dimensional, continuous valued labels. This consensus annotation data will be used in later chapters of this thesis as the basis for an automatic system.

Currently, the cultural background of the NVC encoders and the annotators is not well controlled. To understand and distinguish between personal and cultural differences, it would be beneficial to have recordings from distinct cultures and have them annotated by multiple cultures. This use of culturally distinct annotators is discussed in Chapter 6.

The finding that NVC labels often co-occur and do not vary independently raises the question as to redundant information in NVC labels. If NVC can be expressed in a lower dimensional space, the annotation task and automatic recognition problem are both simplified. However, the NVC labels used are not comprehensive and further work is needed to find an NVC annotation system with a broader coverage of meaning.

The corpus was recorded in a laboratory environment. It is likely that more naturalistic data would be obtained by recording data in an environment in which an automatic system is expected to be deployed. Some data sets that attempt this have recently become available, such as the D64 Multimodal Conversational Corpus being recorded in a domestic environment. The next chapter describes an automatic NVC classification system based on the data collected in this chapter.

Computers are useless. They can only give you answers.

Pablo Picasso

4

Automatic Classification of NVC in Spontaneous Conversation

This chapter describes a study of automatic recognition of natural NVC in informal conversations. To find the features and classifier best suited for this task, a comparison between various alternative components of the recognition system is performed. The automatic system uses the naturalistic NVC corpus described in the previous chapter. This thesis only considers visual information, because it is an important mode for NVC. Although the exact modality distribution of each NVC is not well documented in the literature, possibly due to the enormity of such a task, this study focuses on facial behaviour because it is obviously a very significant modality for NVC[1]. However, other non-facial behaviours are thought to play a significant role in the expression of

[1] "In fact, some researchers believe the primary function of the face is to communicate, not to express emotions." [166], p. 9

NVC. Also, it is well known that the non-verbal component of voice contains sufficient information to recognise emotion. It is likely that a hybrid approach using visual and audio information would be an effective approach for NVC recognition.

Because the TwoTalk corpus is used for training, this is one of the first automatic NVC recognition systems to operate on informal conversation [2]. It is also one of the first to recognise human behaviour outside of a task based, role play based or otherwise specialised situation [3]. This chapter only uses clear samples of NVC, while intermediate intensity samples are discarded (this is discussed in detail in Section 4.4). This simplifies the recognition task and makes the work similar to some previous studies which use strong acted emotion, while retaining the naturalistic quality of the data. However, a practical system should also operate on intermediate intensity NVC; this issue is considered in Chapter 7. There are only a few previous studies of automatic NVC recognition, with agreement and disagreement being the most popular. The annotation labels used in this thesis are *agree*, *thinking*, *question* and *understand*.

The primary contributions of this chapter are:

- A study of automatic NVC classification of clear, naturalistic samples in informal conversation. Different approaches to feature extraction and classification are compared.
- Visualisation and analysis of *thinking* NVC in a simplified feature space based on eye movements, to determine the presence of any consistent behavioural patterns.
- An analysis of classification performance for specific types of NVC signals, to find their relative recognition difficulties.
- A comparison between the feature extraction approach and temporally encoding the features using a quadratic function.

[2] The existing AMI corpus contains informal conversation and has been used for automatic recognition but the labels used generally do not address NVC specifically.

[3] Again, some studies have used the informal part of the AMI corpus has been addressed, but often these studies also consider the acted data as they are part of the same data set.

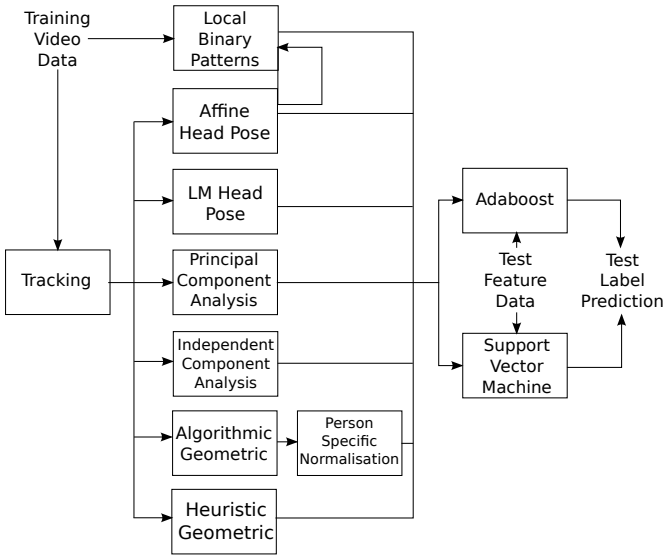

Figure 4.1: An overview of the automatic NVC classification system.

The following section provides an overview of existing classification techniques. Section 4.1 provides a broad overview of the automatic system. Section 4.2 explains the various feature extraction methods used in the comparison. Various feature extraction methods are compared to determine the best approach. Section 4.3 describes the classifiers used in the comparison. The selection of clips for training and testing from the corpus is described in Section 4.4. Performance measurement is detailed in Section 4.5. The performance of various features and classifiers is shown and discussed in Section 4.6. The use of polynomial curves to encode temporal variations is examined in Section 4.7. An exploration of *thinking* in a simplified feature space is conducted in Section 4.8. Work by Akakın and Sankur [10] that builds upon the work of this chapter is discussed in Section 4.9.

4.1 Overview

This chapter describes an automatic NVC classification system and evaluates its performance. The basic steps are shown in Figure 4.1. Corpus data is split into cross validation folds of seen training data and unseen test data. Various feature extraction techniques are used as the baseline for a comparison of different approaches. The fea-

tures are then used to train either an SVM or Adaboost model. The model is used to predict labels on test data, on which the appropriate feature extraction technique has been applied. The predicted labels are compared to annotator ratings and the performance is evaluated.

The next section describes the issues in converting the original videos into suitable features for supervised learning.

4.2 Feature Extraction

This chapter describes several different feature extraction approaches and compares their performance for NVC classification. Feature extraction takes raw input frames and produces a feature representation that is intended to increase the robustness to changes in the raw input data that are not significant for the application. The irrelevant changes include:

- identity and face shape,
- lighting changes,
- head rotation and translation, although it is useful to have this encoded separately from emotion and not entirely discarded and
- occlusions.

Existing feature extraction techniques have previously been discussed in Section 2.4. There are a wide range of possible approaches to facial feature extraction and the choice of features in previous works are largely experimentally driven to achieve accuracy on a chosen data set. However, there are a few properties of naturalistic NVC that make this task distinct from many posed behaviour based studies. Naturalistic behaviour is harder to recognise than posed examples [59], probably due to the differences in the way emotions are expressed in these situations, as well as the amount of head pose present in the data. These large pose changes make some approaches unsuitable for natural NVC recognition. For this reason, AAMs were not used because, in their basic formulation,

they are not robust to large head pose changes [299]. Also, large head pose changes cause significant changes in facial appearance [187] and normalising face appearance to frontal pose is difficult, requiring image completion or texture synthesis to replace self-occluded parts of the face [19]. These factors are likely to make appearance based features less suitable for this task. The effect of large pose changes can be mitigated by the use of a multi-camera recording, but this is a more involved experimental arrangement which limits this approach to fewer practical applications. For these reasons, this study primarily focuses on facial shape recorded using a simple camera. It is assumed that by using the visual modality, only non-verbal communication will be used as the basis for recognition. In principle, it is possible for an automatic system to learn the verbal component of communication via speech reading. However, given the extreme difficulty of speech reading [232], this is not a significant factor in the context of this study. The strengths and weaknesses of the methods employed are discussed in the next few sections.

The names of the feature extraction approaches have been abbreviated for the sake of convenience: *affine* refers to affine head pose estimation, *deform-cubica* refers to tracking deformations based in CubICA, *deform-fastica* refers to tracking deformations based on FastICA, *deform-pca* refers to tracking deformations based on PCA, *geometric-h* refers to heuristic geometric features, *geometric-a* refers to algorithmic geometric features, *lbp* refers to local binary patterns and *lma* refers to head pose estimation by Levenberg–Marquardt Algorithm (LMA) model fitting.

4.2.1 Linear Predictor Feature Tracking for Facial Analysis

LP feature tracking [231] was applied to unconstrained natural conversation videos. This method was selected because it is relatively robust to head pose changes. The theory for this method is discussed in Appendix B and is used instead of the more common trackers, such as Kanade-Lucas-Tomasi (KLT) [306]. Both methods store pixel intensity information for a region near a feature point of interest. However, the KLT feature tracker uses a single training template as a model, while LPs use a model based on one or more training frames. For this reason, an LP tracker requires more

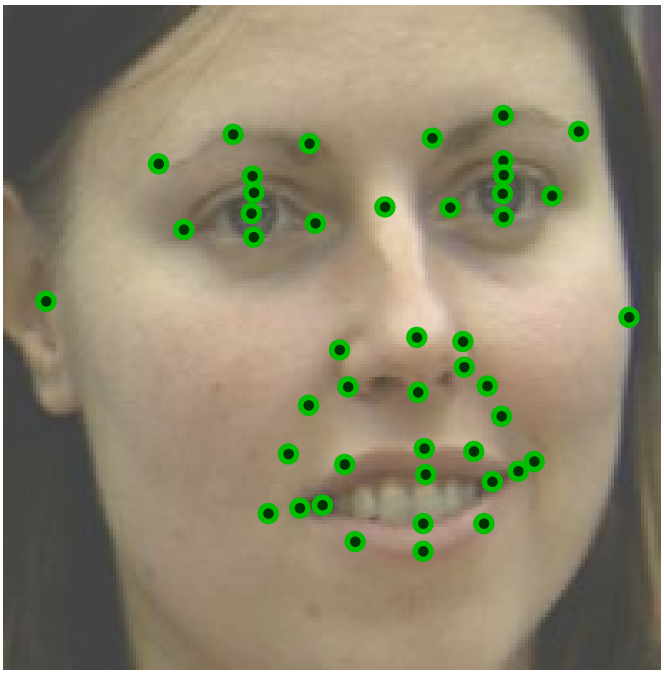

Figure 4.2: Illustration of position of LP Trackers on facial features.

manual interaction by an operator to create suitable training data. However, the use of multiple appearances from multiple frames enables an LP to generalise to multiple poses.

The number and location of the trackers was based on balancing the need to:

- encode as much shape information of the face as possible,
- the need to be able to reliably specify the position on multiple frames (due to multiple frame training being supported by LPs) and
- the resources needed to create the tracker training data.

The κ points on the face that were selected to be tracked are shown in Figure 4.2. The positions were manually initialised at the start of each of the 12 minute videos. The tracker then predicted the feature position on each subsequent frame. Point correspondence was maintained between the eight subjects in the corpus to enable inter-person comparison. Due to extreme head motion or occlusions, the tracker occasionally suffered unacceptable levels of drift and was manually re-positioned on the correct feature

location. The resultant tracking positions for κ trackers on a single frame are designated $\mathbf{T} \in \mathbb{R}^{\kappa \times 2}, \kappa \in \mathbb{N}$

4.2.2 Heuristic Geometric Features

Some feature extraction methods attempt to comprehensively encode the overall face deformation in an unsupervised fashion, but these are not necessarily optimal for classification. An alternative approach is to manually engineer a set of features that correspond to local facial deformations. The relevant areas of the face are manually selected based on experience of what is likely to be relevant information to perform classification. These features are referred to as "heuristic geometric" features, abbreviated to *geometric-h*. The specific feature set used was inspired by el Kaliouby and Robinson's manually engineered features [100] (see Table 4.1 for a comparison). They primarily use shape information from tracking, as well as additional appearance features to find the mouth aperture area and teeth visibility. The approach presented here slightly differs in that it uses shape information only (see Table 4.2). Additional features were added to encode eye gaze. Because the features correspond to muscle driven facial deformations, they are analogous to a subset of FACS action units [95]. Heuristic features are computed based on LP tracker positions (shown in Figure 4.3) to form frame feature $\mathbf{f}_{geometric-h} \in \mathbb{R}^{12}$.

Heuristic features focus on only a subset of facial deformations. The next section describes a method to encoding a broader range of face deformations.

4.2.3 Algorithmic Geometric Features

Manually engineered features encode a subset of facial deformations that are thought to be relevant by their designer. An alternative is to comprehensively encode the shape information of the facial feature trackers. Frame features are exhaustively generated based on a simple geometric measure. This approach is referred to as algorithmic geometric features, abbreviated to *geometric-a*. Ideally, this would encode information pertaining to local deformations of the face in separate features than motion due to head pose changes. Simple distances between pairs of trackers are used (see Figure

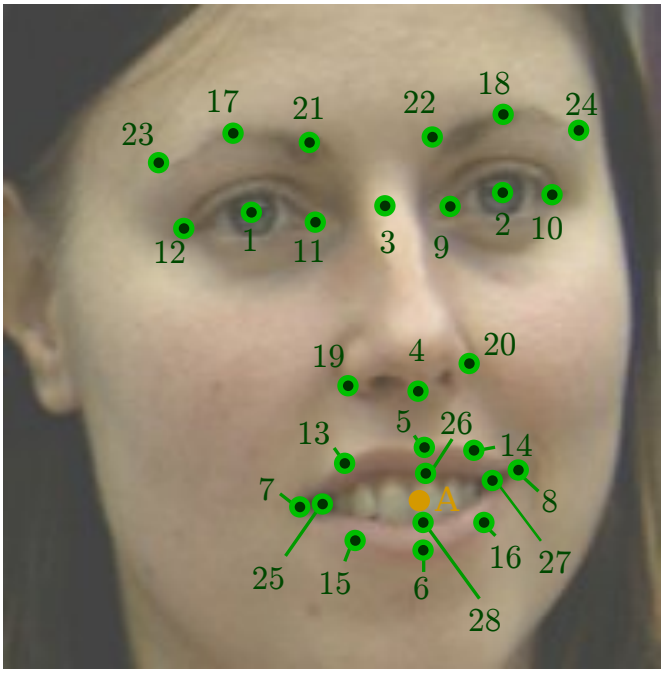

Figure 4.3: Illustration of position of LP Trackers used in the extraction of Heuristic Geometric Features.

Table 4.1: Inter-Culture Correlation of Various Mean Filtered Culture Responses.

Heuristic Features	El Kaliouby and Robinson Features [100]
Head yaw	Head yaw
Head pitch	Head pitch
Head roll	Head roll
Eyebrow raise	Eyebrow raise
Lip pull	Lip pull
Lips part	Lip pucker
Eye horizontal position (Right)	Lips part
Eye horizontal position (Left)	Jaw drop
Eye horizontal position (Mean)	Teeth visible
Eye vertical position (Right)	
Eye vertical position (Left)	
Eye vertical position (Mean)	

Table 4.2: Heuristic geometric features used to extract facial expression while being robust to pose. Position A is the average position of the outer mouth trackers. t is the current frame number. These features were inspired by el Kaliouby and Robinson [100].

Head yaw	$\frac{\overline{\mathbf{T}_9\mathbf{T}_{10}}}{\overline{\mathbf{T}_{11}\mathbf{T}_{12}}}$
Head pitch	$\mathbf{T}_4[t] - \mathbf{T}_4[t-1]$
Head roll	$\angle\mathbf{T}_9\mathbf{T}_{11}$
Eyebrow raise	$\frac{(\overline{\mathbf{T}_{11}\mathbf{T}_{21}}+\overline{\mathbf{T}_1\mathbf{T}_{17}}+\overline{\mathbf{T}_{12}\mathbf{T}_{23}})_t}{(\overline{\mathbf{T}_{11}\mathbf{T}_{21}}+\overline{\mathbf{T}_1\mathbf{T}_{17}}+\overline{\mathbf{T}_{12}\mathbf{T}_{23}})_0}$
Lip pull/pucker	$\frac{(\overline{A\mathbf{T}_7}+\overline{A\mathbf{T}_8})_t-(\overline{A\mathbf{T}_7}+\overline{A\mathbf{T}_8})_0}{(\overline{A\mathbf{T}_7}+\overline{A\mathbf{T}_8})_0}$
Lips part	$\overline{\mathbf{T}_{26}\mathbf{T}_{28}} \cdot \overline{\mathbf{T}_{25}\mathbf{T}_{27}}$
Right eye horizontal	$\frac{\mathbf{T}_{11}\mathbf{T}_{12}\cdot\mathbf{T}_{11}\mathbf{T}_1}{\lvert\mathbf{T}_{11}\mathbf{T}_{12}\rvert}$
Left eye horizontal	$\frac{\mathbf{T}_9\mathbf{T}_{10}\cdot\mathbf{T}_9\mathbf{T}_2}{\lvert\mathbf{T}_9\mathbf{T}_{10}\rvert}$
Mean eye horizontal	$\frac{(\mathbf{T}_{11}\mathbf{T}_{12}\cdot\mathbf{T}_{11}\mathbf{T}_1)(\mathbf{T}_9\mathbf{T}_{10}\cdot\mathbf{T}_9\mathbf{T}_2)}{2\lvert\mathbf{T}_{11}\mathbf{T}_{12}\rvert\lvert\mathbf{T}_9\mathbf{T}_{10}\rvert}$
Right eye vertical	$\frac{\lvert\mathbf{T}_{11}\mathbf{T}_{12}\times\mathbf{T}_{11}\mathbf{T}_1\rvert}{\lvert\mathbf{T}_{11}\mathbf{T}_{12}\rvert}$
Left eye vertical	$\frac{\lvert\mathbf{T}_{11}\mathbf{T}_{12}\times\mathbf{T}_{11}\mathbf{T}_1\rvert}{\lvert\mathbf{T}_{11}\mathbf{T}_{12}\rvert}$
Mean eye vertical	$\frac{\lvert\mathbf{T}_{11}\mathbf{T}_{12}\times\mathbf{T}_{11}\mathbf{T}_1\rvert\lvert\mathbf{T}_{11}\mathbf{T}_{12}\times\mathbf{T}_{11}\mathbf{T}_1\rvert}{2\lvert\mathbf{T}_{11}\mathbf{T}_{12}\rvert\lvert\mathbf{T}_{11}\mathbf{T}_{12}\rvert}$
Mouth centre	$A = \overline{[\mathbf{T}_5, \mathbf{T}_6, \mathbf{T}_7, \mathbf{T}_8, \mathbf{T}_{13}, \mathbf{T}_{14}, \mathbf{T}_{15}, \mathbf{T}_{16}]}$

4.4), in a similar fashion to Valstar et al. [317]. The length of the feature vector $\mathbf{f}_{geometric-a}$, for κ trackers, is triangular number T_κ where $T_\kappa = \frac{\kappa(\kappa+1)}{2}$. In this work, $\kappa = 46$ trackers are used to generate algorithmic features (see Figure 4.2), therefore $\mathbf{f}_{geometric-a} \in \mathbb{R}^{T_{46}} = \mathbb{R}^{1035}$. For two trackers of index $a \in \mathbb{R}$ and $b \in \{0...\kappa\}$, where $1 \leq a < b$, the distance is computed:

$$\mathbf{f}^{a+b+T_{b-1}}_{unnormalised\_alg} = |\mathbf{T}_a - \mathbf{T}_b| \tag{4.1}$$

Although these feature components encode shape information from localised areas of the face, the features tend to include redundant information and are not robust to scale changes.

The features are then processed to reduce the effect of identity. This normalisation is not applied to the other types of features (see Figure 4.1), although later chapters

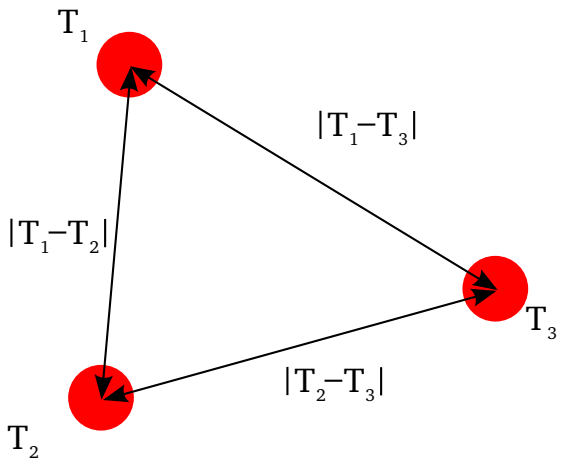

Figure 4.4: A trivial example of how algorithmic geometric features are calculated for 3 tracked points. Exhaustive distance pairs of trackers are calculated.

have normalisation applied in a consistent fashion to avoid this problem. Each feature component is rescaled and zero centred. This removes shape information due to both identity and some types of expressivity by removing person specific differences in facial deformation magnitude. Information that relates to facial shape deformation, in terms of a difference from the average face, is retained. For a video sequence of p frames, all frame feature vectors are concatenated into a frame feature matrix $\mathbf{F} \in \mathbb{R}^{1035 \times p}, p \in \mathbb{N}$. The mean $\mathbf{m}$ and variance $\mathbf{v}$ values of each feature component are used to normalise and zero centre the feature ($\mathbf{v}, \mathbf{m} \in \mathbb{R}^{1035}, i \in \{0...1035\}$, $t \in \{0...p\}$):

$$\mathbf{m}^i = \frac{\sum_{j=0}^{p} \mathbf{F}^{i,j}_{unnormalised\_alg}}{p} \tag{4.2}$$

$$\mathbf{v}^i = \frac{\sum_{j=0}^{p} \mathbf{F}^{i,j}_{unnormalised\_alg}{}^2}{p} - \mathbf{m}^{i\,2} \tag{4.3}$$

$$\mathbf{F}^{i,t}_{geometric-a} = \frac{\mathbf{F}^{i,t}_{unnormalised\_alg} - \mathbf{m}^i}{\sqrt{\mathbf{v}^i}} \tag{4.4}$$

$\mathbf{F}_{geometric-a}$ is easy to compute but it requires an existing set of frames covering the range of NVC signals to calculate the mean $\mathbf{m}$ and scaling factors $\mathbf{v}$ used in normalisation. This makes the approach unsuitable for immediate feature extraction of a

previously unseen face.

4.2.4 Tracking-based Features using PCA and ICA Dimensionality Reduction

PCA and ICA can be used to separate information due to changes in pose from information due to local deformation. This is necessary because raw features from tracking contain translations and other movement information that is not necessarily relevant to NVC classification. Directly using raw features would result in poor recognition performance. These dimensionality reduction techniques subdivide a signal by projecting it on to a new set of basis vectors; the basis vectors in the case of PCA correspond to a particular overall deformation of the face. PCA selects eigenvectors that are the largest orthogonal modes of variation, while ICA selects basis vectors that are statistically independent. The use of PCA on tracking data has previously been used by Lien et al. [178]. The basis eigenvectors are learned in an unsupervised fashion but generally do not correspond to local areas of the face, nor guarantee that they are optimal for recognition.

For κ trackers, the positions on a single frame is designated as $\mathbf{T} \in \mathbb{R}^{2\times\kappa}$. For a video of p frames, the tracking data is reshaped into a $2\kappa \times p$ matrix. This tracking matrix is zero centred to form matrix $\mathbf{M}$. Performing PCA on $\mathbf{M}$ produces 2κ principal components. Each frame was then projected on to these basis vectors to form a frame feature $\mathbf{f}_{pca} \in \mathbb{R}^{2\kappa}$. A similar procedure was used to project frames into ICA space. Two specific ICA implementations were used: FastICA [147] and CuBICA [35]. FastICA is advantageous because of its fast convergence, while CuBICA is able to separate asymmetrically distributed sources but at a higher computational cost [327]. The frame feature corresponding to these methods are referred to as $\mathbf{f}_{fastica}$ and $\mathbf{f}_{cubica}$ respectively.

4.2.5 LMA Head Pose Estimation

The shape deformation of the face has been considered, but appearance information may also contribute to NVC recognition. For example, the appearance of wrinkles or the teeth cannot be effectively tracked due to their transitory appearance. To make

comparisons between different frames and persons, the faces need to be aligned. This enables direct comparison of corresponding positions for the purposes of NVC recognition.

Head pose may be estimated by model fitting and cost minimisation. Cost minimisation is performed by minimising least square errors by Levenberg–Marquardt Algorithm (LMA), and a simple head model based on an average head shape and expressed in homogeneous coordinates $\mathbf{H} \in \mathbb{R}^{\kappa \times 4}$. This is similar to the approach used by Liu and Zhang [183], but the head model is simplified to ignore the effect of expression. The model fitting results in estimates for 3 translation components and 3 Euler rotation components. Because of the simplicity of the model, it cannot encode facial expression but it does encode head pose information in an intuitive way. For example, head nodding and shaking are encoded as two distinct components. For each frame, the head post model error ψ is estimated as follows:

$$\psi(\mathbf{R}, \mathbf{t}) = \sum_{i=1}^{\kappa} ||\mathbf{T}_i - \tau(\mathbf{R}\mathbf{H}_i + \mathbf{t})||^2 \tag{4.5}$$

Where τ is the perspective projection function, $\mathbf{t}$ is the translation (where $\mathbf{R} \in \mathbb{R}^{4\times4}$ is the head rotation matrix corresponding to the Euler angles $\mathbf{R}_{pitch}$, $\mathbf{R}_{roll}$, $\mathbf{R}_{yaw}$ and $\mathbf{t} \in \mathbb{R}^4$ is the head translation $\mathbf{t} = \{\mathbf{t}_x, \mathbf{t}_y, \mathbf{t}_z, 1.\}$). The pose parameters are varied to find the minimum model fit error. The pose variables $\mathbf{R}_{pitch}$, $\mathbf{R}_{roll}$, $\mathbf{R}_{yaw}$ and $\mathbf{t}_x, \mathbf{t}_y, \mathbf{t}_z$ are concatenated to form a frame feature $\mathbf{f}_{lma} \in \mathbb{R}^6$.

4.2.6 Affine Head Pose Estimation

If the tracking positions $\mathbf{T}$ are re-expressed as the homogeneous coordinate matrix $\boldsymbol{\chi} \in \mathbb{R}^{\kappa\times3}$, an affine transform of tracker positions $\boldsymbol{\chi}$ to another frontal reference shape $\boldsymbol{\chi}'$ encodes the head pose. The affine transform can also be used to align the face, but the affine transform approximation breaks down as the face rotates away from the frontal view. The affine transform $\mathbf{A} \in \mathbb{R}^{3\times3}$ can be estimated by taking the matrix Moore-Penrose pseudo-inverse ($\boldsymbol{\chi}^+$) as shown in Equation 4.7. The result can be reshaped to a frame vector vector $\mathbf{f}_{affine} \in \mathbb{R}^6$.

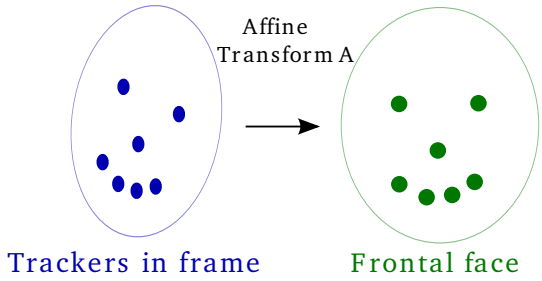

Figure 4.5: An affine transform is calculated to transform the current face on to the frontal face shape. This diagram has been simplified to only use 7 tracked features.

$$\boldsymbol{\chi}' = \mathbf{A} \cdot \boldsymbol{\chi} \tag{4.6}$$

$$\mathbf{A} = \boldsymbol{\chi}' \cdot \boldsymbol{\chi}^{+} \tag{4.7}$$

4.2.7 Uniform Local Binary Patterns

There are many possible approaches to encoding facial texture (see Section 2.4 for background). Local Binary Pattern (LBP) [229] is used in this work, because it has been shown to be effective in encoding facial texture for emotion recognition [285]. LBP focuses on encoding texture information, often in grey-scale images. The role of colour is not considered in this work, because NVC based facial colour changes are relatively rare and are a subtle effect. Local binary patterns are based on comparisons between a central pixel intensity and the intensities of nearby pixels. The comparison is a single "greater than" or "less than or equal" binary choice. The local pixels are often arranged in a simple pattern, such as a circle or a series of concentric circles. In this work, the simplest LBP operator is used, which considers the eight pixels surrounding a central pixel (see Figure 4.6), denoted as $LBP_{(8,1)}$. Each combination of possible binary intensities is mapped into a code book. LBPs in a region of interest are then usually used to form a histogram of code words. Ojala et al. [229] focused on LBPs that have, at most, two bitwise transition in the circle of pixels, which they termed "uniform". These histogram features were found to be effective, while reducing the number of code words from 256 to 59. The histograms are normalised to remove the effect of the number of LBP samples. These histograms may be used as features for

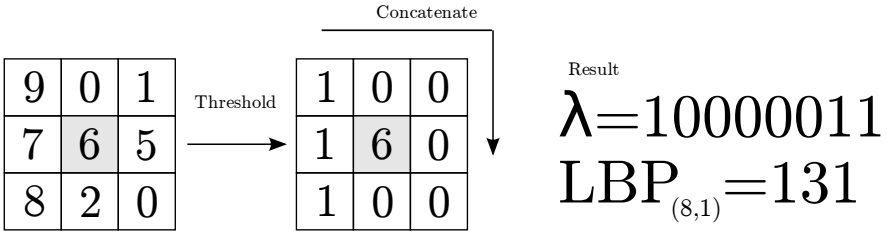

Figure 4.6: The basic $LBP(8,1)$ operator. For each pixel, the adjacent pixels f_i are thresholded λ and concatenated to form an LBP code.

recognition. LBPs are computationally simple and are largely robust to illumination changes, because of the removal of the absolute difference intensity information. For the image intensity of eight pixels $f_i \in \mathbb{R}^8, i = \{0...7\}$ surround central pixel $f_c \in \mathbb{R}$, the pixel comparison vector $\lambda \in \mathbb{Z}^8$ and LBP value $LBP_{(8,1)} \in \mathbb{Z}$ is computed as:

$$\lambda(f_i - f_c) = \begin{cases} 1, & f_i \geq f_c \\ 0, & otherwise \end{cases} \tag{4.8}$$

$$LBP_{(8,1)} = \sum_{j=0}^{7} \lambda(f_j - f_c)2^P \tag{4.9}$$

The face region was subdivided into a grid of g by h rectangles (similar to Feng et al. [114], $g, h \in \mathbb{N}$). The grid used the affine transform $\mathbf{A}$, described in the previous section, to maintain the alignment with the underlying facial features. Uniform LBPs were calculated on each rectangle in the grid, producing $g \times h$ histograms. These histograms were concatenated into a frame feature vector $\mathbf{f}_{lbp} \in \mathbb{R}^{59 \cdot g \cdot h}$.

4.3 Classification for NVC

The automatic NVC system was tested using both Adaboost and SVM classifiers. These classifiers have been shown to be effective in various facial analysis applications. They are both binary classifiers, which are suitable for our constrained problem. The method for comparing the classifier predictions to ground truth is described in Section 4.5. The NVC corpus contains video clips of different lengths but temporal variations are not

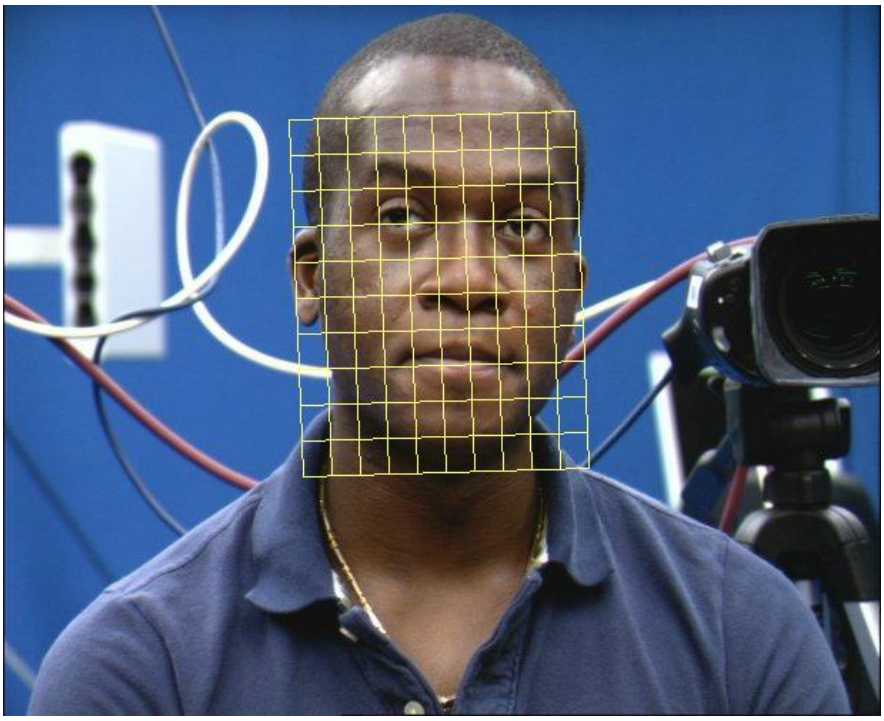

Figure 4.7: Histograms of LBP value frequencies are calculated within each area of a $g \times h$ grid. The grid is aligned to the face using an affine transform which reduces the effect of head pose and translation.

directly modelled by these classifiers. The approach used here is to classify each frame and then fuse the classifier outputs to produce a final label. Basic temporal models that encode information from multiple frames are used in Sections 4.7 and 7.3.1. The training samples used to create the classifier model corresponding to individual frames, however not every frame necessarily contains relevant information. A training clip is used in its entirety, which can lead to a proportion of irrelevant frames being included in the classification model, resulting in a drop in performance. For a clip of length r frames, the s-dimensional frame features $\mathbf{f}$ are concatenated into a clip feature matrix $\mathbf{B} \in \mathbb{R}^{r \times s}$. All clip feature matrices $\mathbf{B}$ are concatenated into a global feature vector $\mathbf{G}$, which for a corpus of ϵ frames: $\mathbf{G} \in \mathbb{R}^{\epsilon \times s}$.

4.3.1 Adaboost Classifier

Adaboost is a supervised binary classifier based on a weighted combination of an ensemble of "weak learner" binary inputs [118]. The input data can contain significant feature noise, but as long as the features are better than random, they are combined by Adaboost to produce a strong classifier. The algorithm operates by incrementally adding feature components to a bank with a corresponding weight. The algorithm selects feature components in an attempt to reduce the training error while focusing on samples that are hardest to classify. At termination, the selected features and weights specify a strong classifier that may be used to predict unseen examples.

Adaboost is simple to implement, computationally efficient, generally avoids overtraining and provides explicit information as to which features are relevant. However, it does not perform as well as other machine learning techniques in some situations. Mislabelled training data can be problematic for some types of classifier, such as Adaboost [220].

All the feature extraction techniques described in Section 4.2 result in continuous value features. However, Adaboost is limited to binary input data and two class problems. For each continuous value feature component, the discretisation is performed by q binary thresholds, designated as $\mathbf{O} \in \mathbb{R}^{q \times s}$. The placement of thresholds for feature component $j \in \{1...s\}$ is determined as follows ($i \in \{1...q\}$):

$$\mathbf{O}_{1,j} = \overline{\mathbf{G}_j} - \sigma(\mathbf{G}_j) \tag{4.10}$$

$$\mathbf{O}_{q,j} = \overline{\mathbf{G}_j} + \sigma(\mathbf{G}_j) \tag{4.11}$$

$$\mathbf{O}_{i,j} = \frac{i-1}{q-1}(\mathbf{O}_{q,j} - \mathbf{O}_{1,j}) + \mathbf{O}_{1,j} \tag{4.12}$$

where $\mathbf{G}_j$ is the jth row of the global feature matrix $\mathbf{G}$ and $\overline{\mathbf{G}_j}$ is the mean of $\mathbf{G}_j$ and $\sigma(\mathbf{G}_j)$ is the variance of the jth feature component. A scaling factor of one standard deviation was experimentally determined. The thresholds are computed on all video samples, which arguably violates the separation of training and test data (SVMs, presented in the next section, do not have this issue.) However, this effect should

be minimal because little person specific information is used. These thresholds are then applied to the feature vector to produce a discretised feature vector $\mathbf{Q} \in \mathbb{R}^{r \times s\ q}$, $a \in 1...r$:

$$\mathbf{Q}_{a,q(j-1)+i} = \begin{cases} +1, & \text{if } \mathbf{B}_{a,i} \geq \mathbf{O}_{j,i} \\ -1, & \text{if } \mathbf{B}_{a,i} < \mathbf{O}_{j,i} \end{cases} \tag{4.13}$$

These thresholds effectively subdivide the feature space by axis parallel hyperplanes with the middle hyperplane positioned on the mean value.

4.3.2 Support Vector Machines Classifier

An SVM is a supervised learning method originally formulated for binary classification [62, 318], although an extension to regression [86] also exists. SVMs use the concept that, although in an original feature space the training samples may not be linearly separable, there exists a non-linear mapping to another space in which a problem is linearly separable. The space is remapped by the use of kernels centred on training samples. An unseen test sample is transformed into this new space and classified based on a simple threshold. The algorithm is difficult to implement efficiently but allows continuous value input variables. This avoids the need to discretise our input features. Unfortunately, the algorithm provides no direct way to examine which features are relevant. This chapter uses the original "C-SVM" formulation of SVMs, rather than the later ν-SVM variant [278]. An SVM may be trained with various kernels; in this study the RBF kernel is used, which is often seen to be effective [285]. A regression variant of SVM, called ν-SVR, is used in Chapter 7.

4.4 NVC Labels and Use of Clear Examples in Classification

The annotation questionnaire was based on four independently varying NVC signals. The four components of NVC rating categories are four independent problems to be

solved. Predictions that distinguish between strong and weak intensity signals, rather than simply positive and negative classification, makes the prediction labels richer and possibly more useful for real applications. However, many machine learning techniques only address classification problems. Also, given the expected difficulty in completely solving the NVC recognition problem (see Section 1.2), this chapter addresses a simpler problem by reducing it to a classification task. This is similar to existing studies conducted on emotion recognition that treated the task as a two class [264] or multi-class problem [57]. The problem of directly recognizing different NVC intensities is addressed in Chapter 7. A set containing all clips in the corpus is designated as V. As discussed in Section 3.6, the consensus mean rating of a clip is denoted $\mathbf{C}$ and contains 4 components corresponding to the four NVC signals.

In this study, clips that were rated as strongly showing an NVC signal were assigned to first positive set and examples that had been rated (by consensus $\mathbf{C}$) as an absence of an NVC signal were assigned to the negative set. Only the 25 highest and 25 lowest ratings were considered as clear examples, as this was judged to be enough for training while excluding more difficult ambiguous examples from both training and testing. Because each component of $\mathbf{C}$ can vary independently, the clip sets containing positive and negative examples of *thinking* are different to the positive and negative sets for *agree*. The clear examples for each NVC category are designated as follows: positive *thinking* set $E^{+}_{thinking}$, negative *thinking* set $E^{-}_{thinking}$, positive *agree* set E^{+}_{agree}, negative *agree* set E^{-}_{agree}, etc. ($E^{+}_{c} \in V, E^{-}_{c} \in V$) The clips are ordered based on the mean annotator rating, $c \in N$:

$$\overline{V}_c = \{i \in \{1, ..., o\} : \mathbf{C}_{c,i} \leq \mathbf{C}_{c,i+1}\} \tag{4.14}$$

The indices of the 25 most positive and 25 most negative clips are identified:

$$\overline{V}^{+}_{c} = (\overline{V}_{c,i})^{o}_{i=o-25} \tag{4.15}$$

$$\overline{V}^{-}_{c} = (\overline{V}_{c,i})^{25}_{i=1} \tag{4.16}$$

The final positive and negative sets are then established:

$$E_c^+ = \{V_i : i \in \overline{V}_c^+\} \tag{4.17}$$

$$E_c^- = \{V_i : i \in \overline{V}_c^-\} \tag{4.18}$$

For a single NVC signal category c, the union between positive and negative sets is then determined for each NVC signal $E_c^{clear} = E_c^+ \cup E_c^-$. There were only a few examples of NVC disagreement and the three other NVC signals had rating scales from neutral to intense expression. For these reasons, examples of disagreement were discarded and *agree* samples were drawn from neutral and positive samples. The next section describes how these machine learning methods are evaluated and compared.

4.5 Performance Evaluation Methods for Variable Length Video Clips

Given the relative difficulty in collecting and annotating data, the available data should be used as efficiently as possible. Therefore, cross validation testing is used, which tests an automatic system in multiple folds. Each fold uses a different partitioning of the data into sets of training and test samples. Eight fold cross validation is used in both person dependent and person independent tests. For person independent testing, this is equivalent to "leave one subject out" testing.

The sample videos have various lengths and the automatic system needs to process them in a way that enables comparison with the true label. The difference between the predicted labels and actual labels is then quantified. Perhaps the most widely used binary classification metrics are accuracy, F1 score and ROC AUC. Apart from being a popular metric, ROC analysis is used because the application and acceptable rate of failure are unknown. ROC analysis shows the system behaviour under a range of false positive rates. The equations for the commonly used metrics are (adapted from [293]):

$$positive = true\_positive + false\_negative \tag{4.19}$$

$$negative = true\_negative + false\_positive \tag{4.20}$$

$$accuracy = \frac{true\_positive + true\_negative}{positive + negative} \tag{4.21}$$

$$precision = \frac{true\_positive}{true\_positive + false\_positive} \tag{4.22}$$

$$recall = \frac{true\_positive}{positive} \tag{4.23}$$

$$f1\_score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{4.24}$$

Computing the AUC of ROC requires a sweep of a decision threshold to determine the false positive (fpr) and true positive rates (tpr). This is often intuitively understood as the area under a plot of tpr vs. fpr, but can also be expressed mathematically as:

$$fpr = \frac{false\_positive(threshold)}{negative} \tag{4.25}$$

$$tpr = \frac{true\_positive(threshold)}{positive} \tag{4.26}$$

$$area\_under\_roc = \int_0^1 trp(fpr)\,\mathrm{d}fpr \tag{4.27}$$

Given a two class problem, there are a few different approaches to process a variable length clip to enable comparison with the true label:

- For each frame in a test clip, the classifier makes a binary prediction. The proportion of positive predictions is taken as the overall positive confidence score. This method is referred to as "clip level" testing. This classification of individual video clips is distinct from event recognition, which is the detection of limited duration events in a longer video [155].
- For each frame, calculate the confidence that the frame is positive. The predictions from multiple frames forms a set of predictions. This set is converted to multiple sets of binary predictions using a moving threshold, as done with a standard ROC analysis. Each thresholded binary prediction is then compared to the

video clip label. This side steps the need for fusion and is referred to as "frame level" testing.

- Combine the frames to form an overall clip feature vector of fixed length. This concept is explored in Section 7.3.
- Use a machine learning method that is dedicated to sequence classification or MIL classification. This is discussed further in Section 4.9.

The first two approaches are used to evaluate performance. They were selected because they both utilise confidence ratings which may be evaluated by an ROC curve.

4.6 Results and Discussion

The feature extraction approaches described above were compared for the four NVC categories. The comparison also includes two machine learning techniques (Adaboost and SVM) and two ways of measuring the performance (frame level and clip level testing). Comprehensive testing was conducted and full results are reported in Appendix A. This section presents the summarised performance results. Various parameters were tuned through experimental validation: five thresholds were used to discretise features for Adaboost ($q = 5$) and an SVM cost parameter $C = 1.0$ was used. LBP grid size were $g = 10, h = 10$.

Table 4.3 shows the results of each of the various feature extraction approaches, as well as comparing multi-person and person independent testing. For brevity, the performance of each of the four NVC categories (*agree*, *thinking*, *understand* and *question*) are averaged to provide a single performance score for each clip. To analyse statistical significance in performance differences, Welch's t-test [330] is employed because the cross fold variances of performance for different methods are unequal. However, this ignores the effect of personal differences in expressivity and style (see Section 2.1), which is likely to cause each cross validation fold to have significant differences in performance. For this reason, the sample variance for this analysis is likely to be inflated and the t-test will be pone to underestimate the true significance. Also, k-fold cross

Table 4.3: Performance of various features and classifiers. Clip level testing, average score of categories shown. SVM with Algorithmic Geometric features produce the highest performance. The error limits are based on one standard deviation of the average cross validation fold performance. The data is shown in graphical form in Figures 4.8 and 4.9

Test	Multi-person		Person independent	
	SVM	Adaboost	SVM	Adaboost
affine	0.52 ± 0.05	0.56 ± 0.04	0.53 ± 0.05	0.51 ± 0.06
deform-cubica	0.50 ± 0.00	0.55 ± 0.06	0.50 ± 0.00	0.51 ± 0.07
deform-fastica	0.50 ± 0.00	0.55 ± 0.06	0.50 ± 0.00	0.51 ± 0.07
deform-pca	0.54 ± 0.04	0.68 ± 0.05	0.50 ± 0.00	0.62 ± 0.07
geometric-h	0.73 ± 0.04	0.68 ± 0.06	0.63 ± 0.05	0.60 ± 0.08
geometric-a	0.75 ± 0.04	0.72 ± 0.04	0.70 ± 0.04	0.68 ± 0.05
lbp	0.58 ± 0.05	0.62 ± 0.05	0.52 ± 0.07	0.48 ± 0.10
lma	0.53 ± 0.06	0.56 ± 0.04	0.49 ± 0.02	0.50 ± 0.08

validation is not an ideal approach to demonstrate statistical significance [80]. The statistical significance analysis in this chapter should be considered in this context. Future work may address this by using additional subjects and a cross validation approach to be more statistically appropriate (see [128]).

Considering person independent SVM classification, *geometric-a* features perform more effectively than *affine* features (significance $p = 0.07$) and *lbp* features ($p = 0.07$). *Geometric-a* features may be more effective than *geometric-h* features, although this effect does not reach statistical significance ($p = 0.26$). For person independent *geometric-a* features, the performance of SVM is not significantly better than Adaboost ($p = 0.43$). For *geometric-a* features with an SVM classifier, person independent classification is not significantly harder than multi-person testing ($p = 0.31$).

Some approaches operate only slightly above chance level (AUC of 0.50) such as head pose features *affine* and *lma*. This suggests that head pose information alone cannot reliably classify NVC signals, although head pose may play a secondary role in NVC expression.

Features based on projecting the face shape information into a ICA space did not result in good performance (*deform-cubica* and *deform-fastica*). PCA based face deformations were more effective than ICA. For the SVM classifier, multi-person testing of *deform-pca* resulted in an intermediate performance of 0.68. When this method was applied to person independent testing, the performance drops to 0.62. This indicates that *deform-pca* SVM creates a model that can predict NVC labels on unseen video clips if the test subject is present in the training data. This pattern of lower person independent performance is repeated for the other feature extraction techniques, when the multi-person and person independent testing performances are compared (see Table 4.3).

The only appearance features used, *lbp*, did not perform as well as other approaches in multi-person testing (0.62 with Adaboost in Table 4.3) despite the use of head pose normalisation. This is surprising because, as discussed in Section 2.4, facial texture is often used for the encoding of facial information. The failure of *lbp* features for NVC signals may be due to one or more of:

- facial texture does not contain information about NVC, which is unlikely,
- person specific face shape and appearance differences reduce the generalisation of classifier models. This should not affect algorithmic geometric features because of the normalisation but features such as LBP will encode person specific information that is not relevant to behaviour recognition,
- an affine transform is used for head pose normalisation. This is a simplistic model and perhaps a more sophisticated model might be effective in removing the effect of head pose, or
- LBPs or *lbp* may be inappropriate for NVC recognition, but a different texture descriptor may be more effective.

Unfortunately, it is difficult to confirm which of these possibilities is true without further experiments.

For multi-person SVM testing features, geometric features (*geometric-h* and *geometric-a*) have the highest performance values of 0.73 and 0.75 respectively (see Table 4.3).

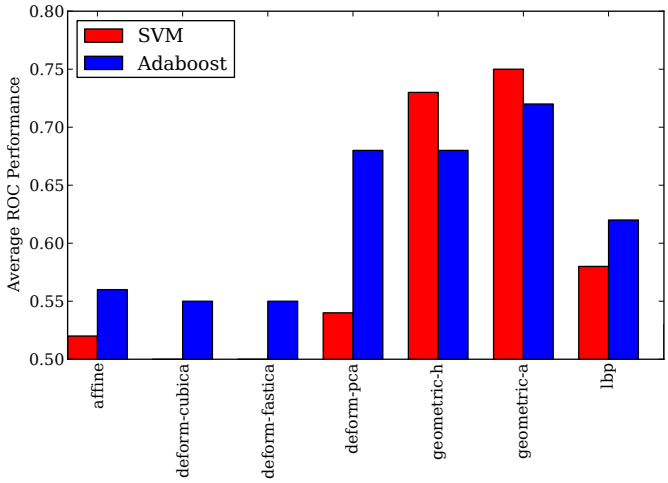

Figure 4.8: Comparison of **multi-person** performance for different features. Testing is at clip level, with the average score of all four NVC categories. A performance of 0.5 is equivalent to classification by chance. SVM with geometric algorithmic features provides the best performance.

However, *geometric-h* features drop to a performance of 0.63 in the person independent case. In the case of *geometric-a*, the drop is less (0.05) to a performance of 0.70, which indicates that these features are less reliant on person specific patterns. Zero centring and scaling features to remove person specific features might benefit facial texture features as well, but this was not performed.

For the higher performing geometric features *geometric-h* and *geometric-a*, SVM performance exceeds the performance for Adaboost. This may be due to information loss in the feature discretisation process, or SVM was better suited to this task.

The performance for each NVC signal is of interest, because some NVCs may be easier or harder to recognise automatically. To limit the quantity of results to a manageable amount, only the top two approaches (SVM with *geometric-a* and *geometric-h*) are presented in the following discussion. Also, person independent testing is used exclusively, since this is a more challenging and general problem.

Table 4.4 shows the *geometric-a* SVM performance for the 4 NVC categories for person independent testing. *Thinking* appears to be the easiest NVC signal to recognise, while *question* is the hardest. Low *question* performance is probably due to the NVC

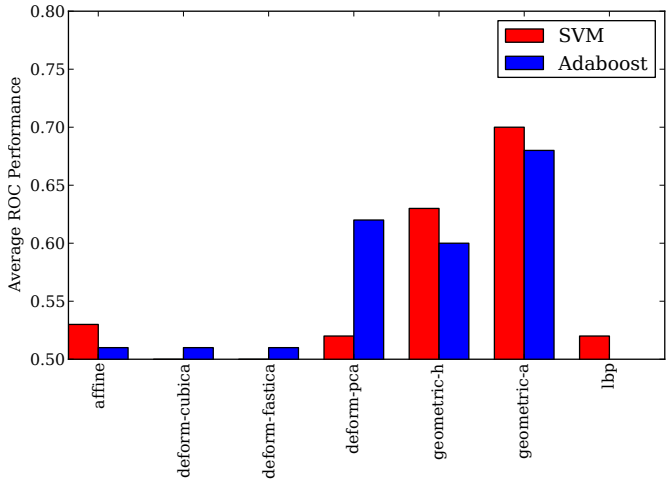

Figure 4.9: Comparison of **person independent** performance for different features. Testing is at clip level, with the average score of all four NVC categories. A performance of 0.5 is equivalent to classification by chance. SVM with geometric algorithmic features provides the best performance.

being primarily expressed by voice intonation [321] and sentence context, rather than any visual cue. Also, *question* category contains fewer positive examples than the other NVC signals in this study (Figure 3.10). Thinking has a characteristic visual appearance which is relatively easy to identify, as discussed in Section 4.8. As previously observed, clip level classification reports a higher performance than frame level classification. However, it is difficult to establish the statistical significance of this result, due to the effect of personal differences increasing the variance of observed performance in the cross validation process. If this effect is ignored and *thinking* NVC is considered,

Table 4.4: Performance for algorithmic geometric features ***geometric-a*** **using SVM** (person independent testing). Confidence intervals of two standard deviations are shown. Classification at the clip level provides better performance then frame level testing (but further tests are required to establish statistical significance). *question* is the hardest NVC category to classify.

Testing	Agree	Question	Thinking	Understand
Clip level	0.67±0.04	0.64±0.10	0.77±0.08	0.71±0.10
Frame level	0.66±0.06	0.54±0.02	0.70±0.06	0.70±0.04

Table 4.5: Performance for heuristic geometric ***geometric-h*** **features using SVM** classification (person independent testing). Confidence intervals of two standard deviations are shown Clip level classification performance exceeds frame level performance. Although performance is generally lower than the use of algorithmic geometric features (Table 4.4), the *agree* NVC performance is significantly better for heuristic geometric features.

Testing	Agree	Question	Thinking	Understand
Clip level	0.73±0.08	0.54±0.16	0.58±0.00	0.68±0.16
Frame level	0.59±0.04	0.51±0.04	0.52±0.04	0.56±0.04

clip level performance (0.77±0.04) exceeds the performance of frame level classification (0.70±0.03) with a significance of only $p = 0.23$. Further experiments are required to establish if this result is statistically significant.

Moving from *geometric-a* to *geometric-h* features, which are shown in Table 4.5, the classification performance is significantly lower overall than in Table 4.4. However, the performance for *thinking* in the case of *geometric-h* exceeds features generated by *geometric-a*. This shows that although *geometric-a* is generally a good method for encoding different NVC signals, it is not necessarily optimal for all types of NVC. As above, statistical significance is low due to the effect of personal differences and further tests are required to establish significance.

4.7 Statistical Features using Quadratic Curve Fitting

Humans use the face shape and appearance variation in time for recognition of behaviour. Temporal variation of features should be investigated to attempt to achieve better automatic performance of emotion and NVC. This information is encoded by statistical features, which are defined here as the result of combining data from multiple sensor observations taken at a range of times. The temporal order of observations may be retained in the feature extraction process, or it may be discarded. To create statistical features, each component of a clip feature (e.g. $\mathbf{B}_{geometric-h}$) is considered independently and in a sliding window (see Figure 4.10). A quadratic curve is fit-

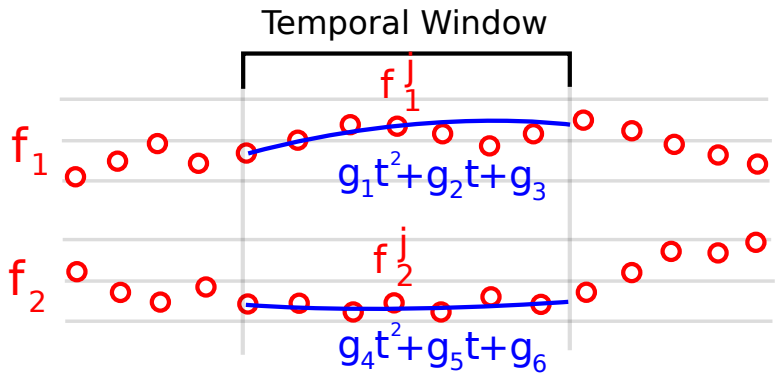

Figure 4.10: Illustration of statistical features for a simple two component frame feature with a single temporal window size. A quadratic curve is fitted to samples in a sliding window. The parameters that describe the curve form part of the temporal vector.

ted to the feature values in the sliding window $\mathbf{W}$ of k frames using least squares fitting $\mathbf{W} \in \mathbb{R}^{k \times s \times r}$. The parameters of the curve $\mathbf{J}$ are then used as the statistical feature, which describes how a frame based feature varies over time in a temporal window $\mathbf{J} \in \mathbb{R}^{r \times 3s}$. This approach is related to Savitzky–Golay filtering [275] and was previously used by Petridis and Pantic [246] to create statistical features for laughter detection. For the ith feature component $i \in \{1...s\}$, at temporal window frame position j, $t \in \{1...k\}, j \in \{1...r\}, \mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$:

$$\mathbf{W}_{t,i,j} = \mathbf{B}_{t+j,i} \tag{4.28}$$

$$P(\mathbf{a}, t) = \mathbf{a}_1 t^2 + \mathbf{a}_2 t + \mathbf{a}_3 \tag{4.29}$$

$$\{\mathbf{J}_{j,3i-2}, \mathbf{J}_{j,3i-1}, \mathbf{J}_{j,3i}\} = \underset{\mathbf{a}}{\operatorname{argmin}} \sum_{t=1}^{k} |\mathbf{W}_{t,i,j} - P(\mathbf{a}, t)| \tag{4.30}$$

Because the optimal size of the temporal window is unknown, multiple window sizes are used to form the statistical features. The number of temporal windows is denoted as e. The combined statistical feature $\mathbf{S}$ is the concatenation of the various temporal windows, and the original frame feature $\mathbf{f}$ on frame j:

Table 4.6: Comparison of AUC performance of statistical features generated based on $\mathbf{S}_{geometric-h}$. SVM Classification was assessed by person independent testing. Confidence intervals of two standard deviations are shown

Statistical Features	Testing	Agree	Question	Think	Understand
No	Clip level	0.73±0.08	0.54±0.16	0.58±0.00	0.68±0.16
Yes	Clip level	0.74±0.06	0.57±0.10	0.60±0.04	0.69±0.12
No	Frame level	0.59±0.04	0.51±0.04	0.52±0.04	0.56±0.04
Yes	Frame level	0.60±0.04	0.51±0.04	0.55±0.04	0.58±0.04

$$\mathbf{S}_j = \{\mathbf{f}_1 ... \mathbf{f}_s, \mathbf{J}^1_{j,1} ... \mathbf{J}^1_{j,3s}, ..., \mathbf{J}^e_{j,1} ... \mathbf{J}^e_{j,3s}\} \tag{4.31}$$

$$\mathbf{S} \in \mathbb{R}^{r \times (3e+1)s} \tag{4.32}$$

The next section discusses their performance in NVC classification.

4.7.1 Results and Discussion

The temporal window lengths used were 80ms, 160ms, 320ms and 640ms ($e = 4$). As before, five thresholds ($q = 5$) were used for Adaboost classification. The SVM cost parameter C of 1.0 was found to be effective.

The performance of statistical features is shown in Table 4.6. Heuristic features $\mathbf{S}_{geometric-h}$ were used instead of $\mathbf{S}_{geometric-a}$ features because its feature matrix is extremely large and exceeds available computer memory resources. The usage of statistical features only results in a slight or negligible improvement in performance in both clip and frame level testing, however further tests are required to establish statistical significance. The performance improvement may be due to input smoothing rather than the linear or quadratic terms in the polynomial being useful. This possibility is supported by feature weights assigned in an Adaboost strong classifier; quadratic and linear terms are generally not selected. In the work by Petridis and Pantic [246], tests showed performance was not significantly affected by varying the temporal window size (Table 2 in

Figure 4.11: Frames from the top 4 annotator rated examples of positive *thinking*. Averted gaze is strongly expressed in positive examples of *thinking*. The specific frames from the clip were manually selected.

their paper). However, humans require temporal information to identify complex emotion, so it is unlikely that temporal information has no role in NVC. Other approaches to statistical features and classification are investigated in Sections 4.9 and 7.3.1.

4.8 Visualising Gaze Features during Thinking

This section will consider a simplified problem that is intended to provide insight into how NVC signals are manifested in feature space. This can inform the decision on how to approach the problem. The NVC for *thinking* is used, because clear positive and negative examples of this signal have a distinctive gaze pattern. This has previously been observed in various studies, such as McCarthy et al. [206], which found that eye

Figure 4.12: Frames from the top 4 annotator rated examples of negative *thinking*, i.e. *thinking* is not present. Eye contact is maintained in negative examples of *thinking*. The specific frames from the clip were manually selected.

contact was broken when a person is thinking about how to answer a question (see Figures 4.11 and 4.12).

Two features from heuristic geometric features, introduced in Section 4.2.2, are used to encode eye movements. Using the full *geometric-h* feature vector would be difficult to visualise as a 2D plot, so the 2 features that are relevant are manually selected. The 9th and 12th feature of *geometric-h* corresponds to "Mean eye horizontal" and "Mean eye vertical" positions respectively (see Table 4.2), which will be referred to as the "gaze subset". These simple features can be used to visualised positive and negative samples.

Each annotated example is a video clip containing multiple frames. The gaze subset of *geometric-h* encode each frame as a 2 component vector. This trajectory is shown

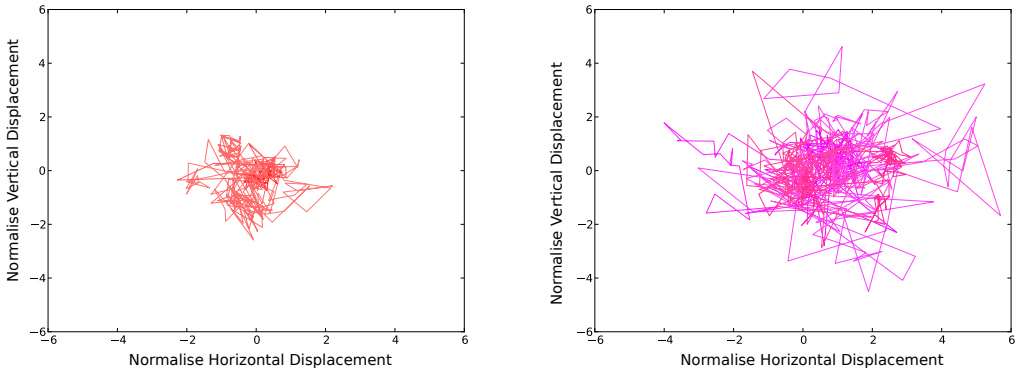

Figure 4.13: Eye trajectories in the top 1% positive and top 1% negative examples of *thinking*. The left plot shows the negative examples. The right plot shows the positive examples. The trajectory shape in the plot correspond to observed eye motion i.e. the top of the plot corresponds to looking upwards. Differences in line shading correspond to different video samples.

in Figure 4.13 as a 2D plot. As can be seen, gaze in negative examples is relatively steady and is generally near the origin, which corresponds to frontal gaze. Any motion away from the origin is due to eye motion or tracker noise. Gaze in positive examples of thinking contain significantly more variation and excursions from the origin to the upper right in the plot. This corresponds to the characteristic thinking behaviour seen in Figure 4.11. This makes positive and negative examples quite distinct in feature space. These patterns in eye movement are likely to be used for automatic *thinking* recognition (this is verified in Section 8.3).

While positive and negative examples are easy to distinguish, intermediate strength examples of *thinking* may prove challenging, if the gaze behaviour is not distinct in gaze feature space. The trajectories of the middle 1% intensity of *thinking* is shown in Figure 4.14. As can be seen, the trajectories are not significantly different from negative samples shown in the left plot of Figure 4.13. This lack of difference may make differentiating between intermediate and negative examples problematic.

The feature can be further simplified by computing the average magnitude of eye deviation from frontal gaze. This also enables representation of a clip by a single number $u \in \mathbb{R}$:

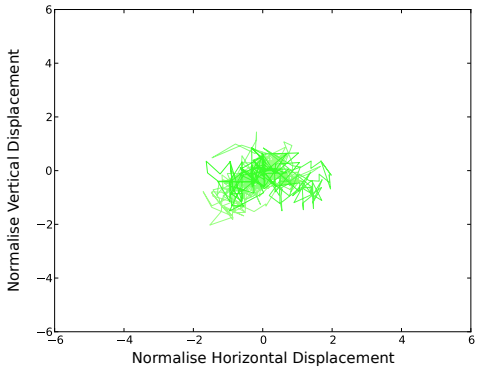

Figure 4.14: Eye trajectories in the middle 1% examples of *thinking*. The trajectory shape in the plot correspond to observed eye motion i.e. the top of the plot corresponds to looking upwards. Differences in line shading correspond to different video samples.

$$u = \frac{\sum_{j=1}^{r} \sqrt{\mathbf{B}_{geometric-h,j,9}{}^2 + \mathbf{B}_{geometric-h,j,12}{}^2}}{r} \tag{4.33}$$

Figures 4.13 and 4.14 only show the trajectories for a small subset of samples. With the mean clip feature u, the feature can be plotted against the annotator rating for all samples in the corpus (Figure 4.15). As can be seen, there is only a weak linear association between this gaze subset feature and the annotators score. The correlation of the feature with the annotation is 0.21. This illustrates the difficulty in NVC recognition: there is little consistency in human behaviour and while particular facial actions may have statistical connections with NVC, they are not definitive. Section 7.3 returns to the concept of summarising an entire clip by a single vector. However, it is important to consider this work in the context of the extreme difficulty of the problem.

4.9 Discussion of Akakın and Sankur, 2011

An early version of the work in this chapter was published in 2009 [287]. A later paper by Akakın and Sankur [10] used the TwoTalk corpus, as well as the BUHMAP database [12] to compare approaches to automatic recognition. Because of the high relevance

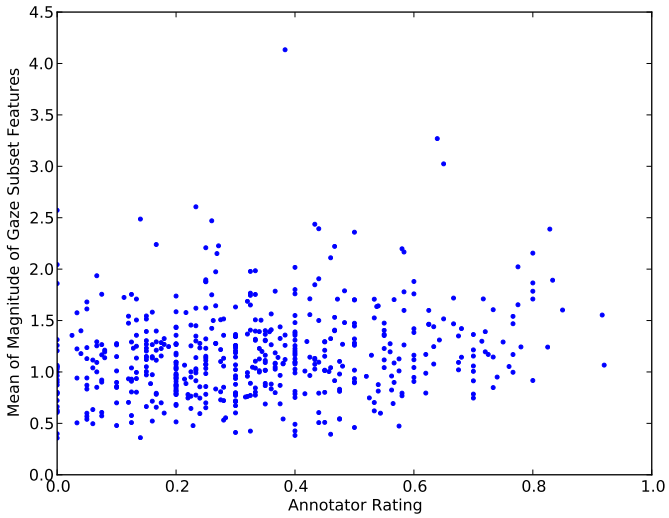

Figure 4.15: The mean magnitude of the gaze features for each clip u, plotted against the annotator rating of *thinking*. There is a weak linear trend between increasing feature magnitude and the intensity of *thinking*.

of their work to this thesis, this section outlines their approach, and discusses their conclusions in light of more recent work.

Their tracking is based on a detection-track-regularise framework with multiple tracking models to account for head pose changes. They used 1208 training frames, which has a higher manual training requirement than LP tracking. Their approach is likely to have better recovery after occlusion and less manual intervention when in operation. The tracking is then used to generate features: landmark coordinate features, heuristic geometric features, and texture features generated from localised patches. These broad types of features are the same as used in Section 4.2.

They use two broad types of machine learning: feature-sequence classifiers (HMM, HCRF) and feature subspace learning (ICA or NMF, classified by Modified Nearest Neighbour (MNN)). A key difference between these methods is feature-sequence classifiers consider the frame order as significant, while feature subspace learning does not. The use of MNN is interesting because it allows direct comparison of an entire trajectory with another trajectory. The authors provide a justification for using feature-sequence

Table 4.7: AUC performance on the TwoTalk corpus, expressed as percentages. The first four rows are from Table 11 in Akakın and Sankur, 2011 [10] and are quoted verbatim. Testing is multi-person. Results from *geometric-a* with an SVM classifier have been appended for comparison (multi-person, clip level testing) (see Table 4.3 and A.4).

Method	Agree	Question	Thinking	Understand	Average
MNN with ICA (P)	84.6	67.6	83.5	74.3	77.5
MNN with NMF (P)	75.8	60	76.1	77.8	72.4
HCRF (17,G)	78.7	75	73.5	82.5	77.4
Classifier Fusion	85.9	78.2	83.8	83.6	82.9
Fused Features, Adaboost	70	73	81	80	76
geometric-a, SVM	70	70	83	75	75

methods by claiming that "even though head and facial gestures may differ in total duration, each follows a fixed pattern of temporal order". While specific gestures generally have a set pattern, it is possible that a particular NVC signal may be expressed by more than one type of gesture.

The results from Akakın and Sankur are reproduced in Table 4.7. All the results presented in the comparison were based on evaluation with the TwoTalk corpus. This table refers to a hybrid approach which is an early version of the approach described in this chapter (see [287] for details). This hybrid approach is a concatenation of the features $\mathbf{f}_{pca}$, $\mathbf{f}_{geometric-h}$, $\mathbf{f}_{lma}$ and $\mathbf{f}_{affine}$ with polynomial temporal fitting applied (see Section 4.7) with Adaboost. The final row in the table corresponds to the method described in this chapter ($\mathbf{f}_{geometric-a}$ and an SVM classifier). As can be seen, the non-hybrid approaches methods have similar levels of performance. The use of a single type of sequential classifiers (MNN or HCRF) only results in 2 to 3% improvement over the *geometric-a* with the SVM approach.

Comparing the non-fused approaches, the best approach for each NVC category are MNN with ICA for *agree* and *thinking*, and HCRF for *question* and *understand*. MNN is poor for *question* NVC and HCRF is relatively poor for *thinking*. The method proposed in this chapter (*geometric-a*, SVM, see Table 4.3) has an overall performance

that is comparable with these non-fused approaches but is worse for *agree* than the temporal modelling approaches. Further tests would be required to establish statistical significance but it is likely that there is little difference in performance between these methods overall. Differences in specific NVC category performance may be due to:

- Each method uses a different representation of the face shape, with each being effective for recognizing a particular subset of NVC signals.
- Temporal modelling may be beneficial for recognizing motion, such as nodding in *agree* which may account for its advantage in performance.
- A feature extraction approach may be advantageous if a particular NVC cannot be characterised by a consistent series of face shapes.

The absence of performance improvement using temporal modelling is similar Petridis et al. [244], who concluded that while the field has moved towards "advanced data fusion methods relying on dynamic classifiers" for human behaviour recognition, the results from their experiments show that the supposed advantage of dynamic classifiers over static classifiers is "not straightforward and depends on the feature representation [...] and the task at hand". In the context of their study, "HMMs does not seem to be beneficial, since it achieves the same performance as a static model". Also in the FERA2011 Facial Expression Recognition and Analysis Challenge [314], Yang and Bhanu's approach was the most effective in both the person independent and overall evaluation and did not temporally model emotions but rather combined a video frames into an emotional avatar frame [339]. Hybrid approaches that combine dynamic and static classification were not used in this thesis but often have higher performance than non-hybrid methods. Interestingly, based on their BUHMAP based tests, Akakın and Sankur conclude that geometric features are superior to texture based features in both sequential and subspace classifiers. This finding concurs with the results presented in this chapter. Binary classification of clear examples of NVC is re-examined in Section 7.5.

4.10 Conclusion

This chapter is a study of automatic classification for common, clear NVC signals in informal conversations. Various feature extraction approaches are compared, as well as two classifiers. The best performance is achieved by a shape based geometric feature that exhaustively computes distances between pairs of trackers. However the method does not encode temporal information in the classifier model. Temporal encoding of feature variations using polynomial curve fitting was not found to increase performance significantly. Also, person specific normalisation of features was only applied to the algorithmic geometric features, and not to the other feature extraction methods. Person specific normalisation should greatly benefit appearance features such as LBP because of the differences in facial appearance across different people can be removed and generalisation can be improved. Later chapters improve on the work here in that they have person specific normalisation applied uniformly to all feature extraction methods.

In this study, shape is found to be more effective than appearance features, which agrees with similar findings in other papers [10] [189], although some studies found appearance based features at least as significant [190] [315].

The differences in cultural background in both the emotional encoders (the people who expressed the NVC) and the annotators (the people who perceived the NVC) is not considered. Cultural differences in NVC perception are revisited in Chapter 6, in which culturally specific annotation data is collected. The next chapter attempts to create an NVC recognition system based on communication backchannel, which loosely speaking is a "NVC listener's" response to a communication event.

The meeting of two personalities is like the contact of two chemical substances: if there is any reaction, both are transformed.

Carl Jung

5

Interpersonal Coordination in NVC

The previous chapter considered automatic NVC recognition based on visual information of the sender or "encoder" of an NVC signal (encoder in this case does not refer to the annotators or automatic encoding but rather to the observed subject exhibiting a behaviour [173]). This chapter considers the behavioural associations between an NVC encoder and the NVC perceiver, and describes a study of recognising NVC signals based on the behaviour of the NVC perceiver. In natural two person conversations, the behaviour of one participant influences the other participant. Usually, people take turns to speak in a conversation. In a dyadic (two person) conversation, the information flow is bi-directional. In a particular speech turn, the speaker's communication is referred to as the "forward" channel. The person listening to the speaker also responds to the speaker, primarily using non-verbal communication, and this communication is known as the "backchannel". Backchannel communication allows a listener to influence a conversation without verbally interrupting or taking a conversation turn. The

forward and backchannel communications occur simultaneously and can vary in style across cultures. For this reason, both subjects in a dyadic conversation can be thought of as both an encoder and perceiver of NVC simultaneously and this is bi-directional flow of information exploited in this chapter.

In natural conversations, participants often mimic or mirror behaviours being expressed by the other person. Mimicry or "behaviour matching" occurs when two or more people show similar body configurations [33]. For example, if one person touches their face during a conversation, the other person is more likely to do so. Another class of behaviours is performed in a common rhythm, with the starting and ending of movement being simultaneous. This is known as "synchrony", which comprises of corresponding rhythms, simultaneous movement and the smooth meshing of interaction [33]. Together, synchrony and behaviour matching are considered as "interpersonal coordination". The phenomena of behaviour mirroring and synchrony are not disjunct and can occur at the same time [75]. These effects raises the possibility that a perceiver's behaviour in response to an NVC signal may be useful in NVC recognition. Based on this idea, backchannel information for automatic NVC recognition is studied and evaluated.

The main contributions of this chapter are:

- A method for automatic identification of certain types of interpersonal coordination in the deformation of the face during casual conversation. This method is applied to recorded videos and the results are analysed.

- A study of feature extraction methods for NVC classification that identifies NVC signals observed in the forward channel, based solely on backchannel feature data.

The next section outlines previous studies that are related to this work. Section 5.2 investigates the inter-person coupling of face deformations in conversation. A automatic system to classify NVC based on backchannel cues is described in Section 5.3.

5.1 Related Research

Backchannel is a form of human communication which has been relatively little studied. The first mention of it is probably by Yngve [341] where he observed that the backchannel allows a listener to express non-verbal or brief verbal speech to influence a conversation, without taking a conversational turn. Even prior to this, gaze was thought to be used in conversation regulation by Kendon [161]. Backchannel signals provide a way for a listener to signal agreement or disagreement, as noted by Bousmalis et al. [37]. Backchannels differ across cultures, with some cultures having a higher frequency of backchannel signals [331].

Human behaviour may also be coupled during conversations, besides consciously performed communication acts. Various studies have tried to quantify this behaviour (see Reidsma et al. [259] and Delaherche [75] for a review). Chartrand and Bargh [52] found that people in conversation tend to adopt the same posture, mannerisms and expression. Similar observations have been found for limb movement, and many aspects of speech, but these effects are dependent on social context [120]. Several studies have used automatic recognition of synchrony (see [75, 297] for reviews). There is a many different feature extraction techniques and measures of synchrony. Richardson et al. [262] studied gaze and noted that when two subjects are observing the same scene, their gaze motion is coupled. They found that behaviour was most strongly coupled between synchronised inter-person movements, but also found evidence of gaze coupling occurring after a delay of up to about 3 seconds. The social situation has an impact on the types of mimicry that occur [36]. Coupled human behaviour may be useful for automatic recognition. Until recently, there have been no approaches that use backchannel communication. Morency [215] studied automatic recognition based on backchannel signals and attempted to predict listener responses to a speaker. Okwechime et al. [230] used a-priori data mining to find behaviour patterns in conversation and considered both audio and visual modalities. This study was performed in controlled social situations to elicit interested or disinterested behaviours by the listeners and social context based behaviour differences were identified. Ramseyer and Tschacher [255] used correlations in difference images to find cross-correlated patterns with a time offset of

±5 sec. They used time shuffled windows to produce non-coordinated behaviour data to determine the extent of cross correlation cause by the null hypothesis. No existing studies consider automatic detection of interpersonal coordination in facial behaviour, although Sun et al. are working towards this goal [297].

The MAHNOB HMI iBUG Mimicry dataset (shortened to MHi-Mimicry-db) [296] contains recordings of dyadic conversations in role play and discussion situations for the purposes of analysing mimicry behaviour. This dataset is significantly larger and more complex than the TwoTalk corpus, containing about 12 hours of records of 40 subjects using 15 video cameras and 3 audio channels. The corpus was manually annotated to produce labels to investigate how mimicry is expressed, if it is intentional or unintentional and the social signal being expressed. Intentional signals were also labelled in terms of the goal of the mimicry. The camera synchronised error was "well below" 20 μsec. The two social situations contained in the MHi-Mimicry-db were a political discussion with a confederate and a role play concerning renting a room to a potential lodger. These scenarios were selected to test specific existing hypothesises of human behaviour. Both were dyadic conversations were conducted in a laboratory. The MAHNOB Mimicry dataset was not available at the time this part of the study was conducted. The following section focuses on interpersonal coordination that may be automatically identified without the use of annotation.

5.2 Coupling in Interpersonal Coordination during Informal Conversation

This section describes an automatic method to analyse coupling in facial behaviour. Because this correlation based method is sensitive to both rhythmic and non-rhythmic behaviours, certain types of both mimicry and synchrony behaviours will be detected. Correlation being used as a measure of mimicry can be understood in terms of it's mirroring property. Two people tend to reflect or mirror each others body positions and behaviours. By analogy, correlation is a way of measuring the extent of "mirroring" between two variables. This makes the Peason's correlation suitable for finding some

types of coupled behaviours. A more detailed justification of using Pearon's correlation is provided in Appendix E.

Correlation of simultaneous frame features is not sensitive to types of mimicry in which the response behaviour is delayed for a longer time than the duration of the trigger behaviour. The class of behaviours that can be detected by the approach described here will be referred to as "coupling". Automatic methods can provide a more comprehensive analysis than can be achieved by manual methods because automatic approaches can scale to large quantities of video data. The next section describes the method, based on tracking, feature extraction and Pearson correlation. Section 5.2.2 shows the results and discusses their significance.

5.2.1 Methodology

Using the four dyadic conversations recorded as part of the corpus (see Section 3.2), behaviour patterns that are common to both participants can be identified. The cameras used to record the corpus are genlocked to ensure synchronisation of video frames. Genlock is a common technique to synchronise cameras using a generator signal. The genlock signal was provided by a Tektronix TG700 Multi-Format Video Generator. The cameras were all of the same type; this greatly simplifies achieving accurate synchronisation. Based on informal tests performed by the broadcast engineer responsible for maintaining the system, the camera jitter measured to be was less than one microsecond (the level of jitter in the camera synchronisation was smaller than measurable using the available equipment). The features introduced in the previous chapter are used, *geometric-a*, which are based on distances of pairs of feature trackers (Section 4.2.3). These features are applied to the face and non-facial behaviour is not considered. Clearly, some areas of the face will not be coupled e.g. when one person speaks, it is usual for the other person to listen and not to move their mouth in the same way. To identify coupled behaviour, the variation of each feature for both speakers in a conversation are compared using Pearson's correlation ρ (see Equation 3.7). If features are highly correlated, this indicates that they are closely coupled. A low correlation indicates the features vary independently. As the correlation is performed

on simultaneous frames, it will only capture mimicry between individuals if the offset between the occurrences is less than its duration. While this approach is insensitive to more delayed mimicry, it should identify shorter term mirroring of behaviour (see Appendix E). The correlation score corresponds to the strength of coupling of the behaviour. The original video records are used because it increases the quantity of data available for analysis. The feature vectors $\mathbf{F}$ of two participants A and B are compared $A, B \in \{1008, 1011, 2008, 2011, 3008, 3011, 6008, 6011\}$. For feature component $i \in \{0...s\}$, the correlation $\mathbf{D}$ is ($\mathbf{F}_i^A \in \mathbb{R}^p$, $p = 12 \times 60 \times 25 frames$, $\mathbf{D}_i^{A,B} \in \mathbb{R}$):

$$\mathbf{D}_i^{A,B} = \rho(\mathbf{F}_i^A, \mathbf{F}_i^B) \tag{5.1}$$

If a behaviour is performed as a reaction, there may be a time delay between the behaviours. Each frame of one participant is compared to a simultaneous frame of the other participant taken using a second camera. This makes this approach sensitive to synchronised behaviours but insensitive to mirroring behaviours that have a temporal offset and short duration. For example, if one person adopts an expression for several seconds and the second person quickly adopts the same expression, this will be found using the correlation of features. The correlation coefficient is limited to measuring linear relationships. It is quite possible that reliable, non-linear patterns of human behaviour may exist. Another limitation of this approach is that this considers only variations in shape but it is possible that patterns in human behaviour may be found if speed or acceleration of the face is also considered.

For this test, corresponding features are compared (e.g. position of left eye in both participants) rather than different areas on the face (e.g. left eye for one subject and mouth opening for the second subject). This constraint makes the system focus primarily on mirroring behaviour. Because the strongly coupled areas of the face are most interesting, the highest scoring feature $i_{max}^{A,B}$, with correlation $\mathbf{D}_{max}^{A,B}$, is found:

$$i_{max}^{A,B} = \underset{i}{\operatorname{argmax}}\, \mathbf{D}_i^{A,B} \tag{5.2}$$

This feature corresponds to a specific area on the face in which the behaviour is coupled. However, there is a possibility that features vary and may coincidentally vary together. It can be difficult to distinguish weak causal relationships with coincidental inter-relationships. To examine this possibility, the correlations of features $\mathbf{D}_{max}$ for individuals that were not in the same conversation was also computed. Apparent correlations between unrelated conversations can be regarded as coincidental.

This method is similar to Ramseyer and Tschacher [255] but contains some important differences:

- This study uses face shape based on tracking rather than difference images of whole people. This makes our approach more effective in localising coupling behaviour to a specific area of the face. However, no other areas of the body are considered in this study.
- This study uses recordings of a different person to test the null hypothesis correlations, rather than a shuffled window approach. However, this should not give rise to a significant difference in measured performance.
- Ramseyer and Tschacher considered time offset signals, which enabled them to see which person is leading and which person is following in behaviour. This is not attempted in this study.

The current section only aims to show that interpersonal coordination exists for some areas of the face in dyadic informal conversation. Therefore, this study consider pairs of participant independently but without attempting to find consistent behaviour patterns that occur in all dyads. This approach of considering dyads independent and without finding corresponding behaviours across dyads was also used by Ramseyer and Tschacher [255]. Although attempting to find consistent behaviours from multiple dyads would be an interesting study, it is beyond the scope of this thesis.

5.2.2 Results and Discussion

The maximum correlation $\mathbf{D}_{max}$ of different pairings of subjects is shown in Table 5.1. The conversation pairs are shown in Table 3.3. There is a weak but significant correla-

Table 5.1: Maximum correlation $\mathbf{D}_{max}$ of corresponding facial shape features for different pairs of participants. Pairs that were participating in the same conversation are highlighted.

	1008	1011	2008	2011	3008	3011	6008	6011
1008	1.00	0.25	0.12	0.11	0.15	0.13	0.08	0.10
1011		1.00	0.19	0.13	0.18	0.10	0.15	0.21
2008			1.00	0.38	0.11	0.16	0.12	0.14
2011				1.00	0.14	0.18	0.13	0.12
3008					1.00	0.34	0.09	0.20
3011						1.00	0.09	0.09
6008							1.00	0.38
6011								1.00

tion (between 0.25 and 0.38) for pairs involved in the same conversation. This confirms our expectation that there are interpersonal coordinated behaviours for corresponding areas of the face. Comparing this to conversation pairings in which the correlation would be coincidental, the correlation is found to be consistently lower than 0.25. This implies that the coupling of facial behaviour in conversation is above the level of correlation due to coincidental matches. Not all highlighted conversation pairs have the same level of correlation. This may be due to different relationships and communication styles between conversation participants. There may be person specific differences in their tendency to couple behaviour, as found by Chartrand and Bargh [52].

The statistical significance of the best correlated features can be confirmed by characterising the distribution of correlation performance scores for the null hypothesis and use it to calculate the p-value for the hypothesis of interest. The possibility of a null hypothesis can be discounted if this p-value is lower than the desired significance level α and can then be considered as statistically significant (see Hayes [138]). Because multiple hypothesises are tested, α is adjusted using the Bonferroni correction to account for the greater chance of finding coincidental patterns. Any correlations between feature pairs that originate from separate conversations are due to chance. The distribution of correlation scores for null hypothesis pairings is shown in Figure 5.1. The

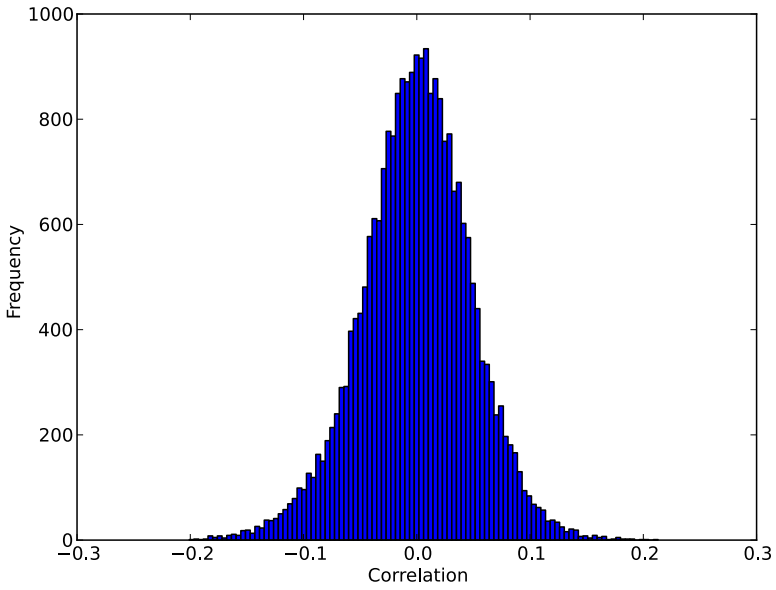

Figure 5.1: Histogram of correlation for corresponding algorithmic geometric features that originate from *different* conversations. The standard deviation is 0.048 and the variance is 0.002.

standard deviation of correlation scores from null hypothesis pairings is $\sigma = 0.048$ and the variance is $\sigma^2 = 0.002$. The chance correlation distribution is zero centred $\bar{\rho} = 0$. The p-value of a correlation score ρ is calculated by a z-test. A z-test is used because the standard deviation of all possible null hypothesis is known. For a single observation compared to a zero centred population $\bar{\rho} = 0$, the z-test score is defined as:

$$z = \frac{\rho}{\sigma} \tag{5.3}$$

For the highest correlation by chance, $\rho = 0.21$, the corresponding z-score is $z = -0.21/0.048 = 4.4$ and p-value is $2 \times \mathcal{N}(4.4) = 1.2 \times 10^{-5}$ where $\mathcal{N}$ is the cumulative normal distribution. The double tailed score is used because the modulus of the correlation is used to determine the largest magnitude correlation. The standard significance level α is usually chosen to be either 0.01 or 0.05. The more stringent level $\alpha = 0.01$ is used in this chapter. Each of the correlation scores use the maximum value, based

Figure 5.2: Corresponding facial distances found to be most coupled in natural conversation for two of the conversations, marked in green. The top row is conversation 1008-1011. The bottom row is conversation 3008-3011.

on 1035 comparisons (see Section 4.2.3). The Bonferroni corrected α is calculated as $\alpha_{adjusted} = \frac{0.01}{1035} = 9.7 \times 10^{-6}$. The p-value of 1.2×10^{-5}, observed in the validation experiments, is within this threshold and cannot be considered as a statistically significant finding, as should be expected.

For pairings in which are engaged in the same conversation, patterns that are clearly statistically significant are expected. The correlation scores for these conversations vary from $\rho = 0.25$ to $\rho = 0.38$. This corresponds to p-values between $2 \times \mathcal{N}(-0.25/0.048) = 1.9 \times 10^{-7}$ and $2 \times \mathcal{N}(-0.38/0.048) = 2.4 \times 10^{-15}$, respectively. These are well below the Bonferroni corrected significance level of 9.7×10^{-6}. The null hypothesis may therefore be rejected for these tests and they are therefore statistically significant.

Figure 5.3: Corresponding facial distances found to be most coupled in natural conversation for two of the conversations, marked in green. The top row is conversation 2008-2011. The bottom row is conversation 6008-6011.

The two areas of the face found to be coupled in natural conversation are shown in Figures 5.2 and 5.3. The areas shown in Figure 5.2 seem to involve vertical distances which are generally rigid and likely encode head pitch. In contrast, the second group's relevant areas, shown in Figure 5.3, are more related to the mouth and perhaps relate to mutual smiling or some other mouth related expression. The next section attempts to find patterns in facial behaviour that are not necessarily related to corresponding areas of the face.

Table 5.2: Maximum correlation $\mathbf{D}_{max}$ of **facial shape features** for different pairs of participants (including corresponding and non-corresponding features). Pairs that were participating in the same conversation are highlighted.

	1008	1011	2008	2011	3008	3011	6008	6011
1008	1.00	0.30	0.18	0.15	0.23	0.22	0.15	0.18
1011		1.00	0.25	0.22	0.29	0.23	0.21	0.23
2008			1.00	0.43	0.17	0.23	0.17	0.18
2011				1.00	0.22	0.19	0.18	0.19
3008					1.00	0.49	0.18	0.29
3011						1.00	0.21	0.16
6008							1.00	0.42
6011								1.00

5.2.3 Coupling of Shape for Non-corresponding Facial Features

The previous section considered the interpersonal coordination of facial behaviour for corresponding areas of the face. However, facial behaviour coupling between people may be expressed in different facial areas. This falls within the definition of synchrony, for which "the important element is the timing, rather than the nature of the behaviours" [75]. Synchrony can involve coordinated times of different forms of behaviour, e.g. beginning to speak when another person stops speaking. This phenomena has been explored to some extent by Ramseyer and Tschacher [255], who did not use corresponding parts of the body but instead searched entire images for any movement synchrony. To find correlations for this broader problem, the feature component i for subject A is compared to feature component j for subject B. Because each feature component corresponds to a different area of the face, this searches for relationships in shape between non-corresponding facial areas. The correlation between facial components i and j ($i \in \{0...s\}$ and $j \in \{0...s\}$) defined as $\mathbf{D}_{i,j}^{A,B} \in \mathbb{R}$ for subject A and B:

$$\mathbf{D}_{i,j}^{A,B} = \rho(\mathbf{F}_i^A, \mathbf{F}_j^B) \tag{5.4}$$

$$(i_{max}, j_{max}) = \underset{i,j}{\operatorname{argmax}} \mathbf{D}_{i,j}^{A,B} \tag{5.5}$$

Again, the components i_{max}, j_{max} with the highest correlation $\mathbf{D}_{i,j}^{A,B}$ are determined. The result of this process is shown in Table 5.2. The correlation for pairs in the same conversation is consistently above the correlation of coincidence matches. The standard deviation of correlation scores from null hypothesis pairings is 0.047 and the variance is $\sigma^2 = 0.002$, which is similar to the case of chance correlations if only corresponding features are considered (see Figure 5.1). More comparisons are performed to obtain the maximum correlation for each pair of subjects. This results in a more stringent Bonferroni adjusted α, which is calculated as $\alpha_{adjusted} = \frac{0.01}{T_{1035-1}} = 1.87 \times 10^{-8}$ where T_i is the ith triangular number. For pairs of subjects in engaged in conversation, the maximum correlation for each pair is 0.30, 0.43, 0.49 and 0.42. Based on the z-test, these correlations have a p-value of 1.7×10^{-10}, 5.75×10^{-20}, 1.90×10^{-25} and 4.03×10^{-19} respectively. The first is relatively near the significance boundary, while the latter three results are clearly statistically significant. There may be because the broader range of facial areas considered includes a wider range of human expression, so instead of just co-occurrence behaviours, other strongly coupled communication relationships can be detected. Because a greater number of feature components are compared, there is an increased possibility for coincidental matches. This is evident in the table because the off axis pairings show an increase in correlation compared to Table 5.1. These results indicate that some facial behaviour is coupled and this coupling is stronger if non-corresponding facial areas are also considered. Although the correlation scores are above chance occurrence, they are not perfectly consistent patterns of human behaviour. This does not make this patterns uninteresting or insignificant because most inter-personal behaviours are only general patterns; identification of deterministic patterns in human behaviour are not the norm.

Table 5.3: Maximum correlation $\mathbf{D}_{max}$ of sliding window **variance of facial shape**, for various pairs of participants (including corresponding and non-corresponding features). Pairs that were participating in the same conversation are highlighted.

	1008	1011	2008	2011	3008	3011	6008	6011
1008	1.00	0.32	0.13	0.13	0.29	0.38	0.16	0.27
1011		1.00	0.33	0.21	0.19	0.16	0.21	0.18
2008			1.00	0.34	0.11	0.23	0.16	0.21
2011				1.00	0.12	0.14	0.26	0.29
3008					1.00	0.41	0.47	0.14
3011						1.00	0.27	0.25
6008							1.00	0.15
6011								1.00

5.2.4 Coupling of Facial Shape Activity for Non-corresponding Facial Features

The previous sections have examined coupling of face shape in natural, dyadic conversation. However, it may be that facial motion contains reliable patterns of human behaviour. For instance, head pitch activity might encode nodding in agreement and this may be related to mouth activity caused by talking. The variance of each feature component in a sliding window was calculated. The window was selected to be 1 second in duration (25 frames), because this is sensitive to NVC signals of a relatively short duration. However, different NVC signals may occur on other time scales which would not be detected. At sliding window position t, the variance $\mathbf{L}$ of a feature component in a sliding window of 25 frames duration is:

$$\mathbf{L}^A_{t,i} = \frac{1}{25}\sum_{a=-12}^{12} {\mathbf{F}^A_{t+a,i}}^2 - \left(\frac{1}{25}\sum_{a=-12}^{12} \mathbf{F}^A_{t+a,i}\right)^2 \tag{5.6}$$

$$\mathbf{D}^{A,B}_{i,j} = \rho(\mathbf{L}^A_i, \mathbf{L}^B_j) \tag{5.7}$$

The results of this method are shown in Figure 5.3. In this case, the correlations caused by coincidence (shown in the off axis pairings) are often higher than the correlations that

might be expected to have coupling. The variance of correlations from null hypothesis pairings has a standard deviation of 0.034. For subject pairs engaged in conversation, the maximum correlations were 0.32, 0.34, 0.41 and 0.15. Based on the z-test, the p-values for these results are 4.88×10^{-21}, 1.52×10^{-23}, 1.74×10^{-33} and 1.03×10^{-05} respectively. The first three pairings are well below the significance threshold $\alpha_{adjusted} = \frac{0.01}{T_{1035-1}} = 1.87 \times 10^{-8}$ but the final pair is above the threshold and is more likely to be due to a null hypothesis. The highest correlation is for conversation 3008-6008 which has $\mathbf{D} = 0.47$ and is caused by coincidental variation of the features (p-value 1.84×10^{-43}). This coincidental match is strangely below the significance threshold and may be due to the distribution of correlations being non-Gaussian. Since coupling cause by patterns in human behaviour cannot be distinguished from those cause by coincidence, more data is required to obtain results that can be confidently considered as statistically significant. An analysis of facial shape activity based on corresponding features resulted in a similar, non-significant result and has been omitted for brevity. This system was not compared to human performance for identifying interpersonal coordination because the resources required to collect such an annotation data set would be very resource intensive and this is beyond the scope of this thesis. Facial shape in natural conversations exhibits interpersonal coordination. Given the relationship between people's behaviours, the next section uses backchannel information for automatic NVC recognition. However, the following section considers all feature components rather than the components identified in earlier in the chapter. This allows features to be used in the classifier model that do not necessarily have a linear coupling with the forward channel behaviour.

5.3 Classification Based on Backchannel Information

The previous chapter has investigated automatic methods to classify NVC based on visual facial information. However, the corpus has two subjects interacting. The annotated subject is designated as the "sender" and the other conversation participant as the "receiver". The sender expresses in the forward channel and perceives the backchannel. The receiver expresses in the backchannel and perceives the forward channel. The pre-

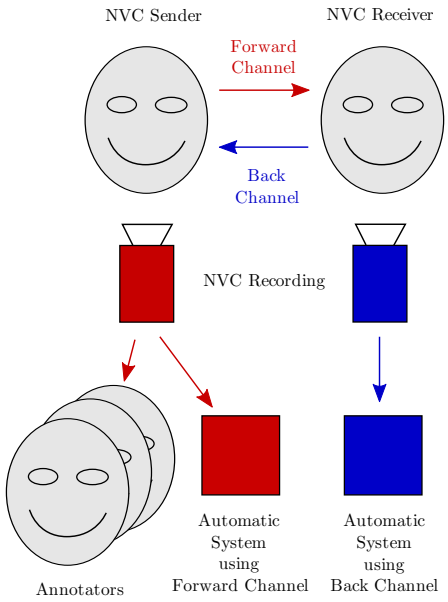

Figure 5.4: Automatic NVC classification can operate on either the forward or backchannel information.

vious section has discussed coupling of facial behaviour in natural conversation. This raises the possibility that a sender's NVC signals may be inferred based on the receiver's visual information. Figure 5.4 shows the experimental arrangement to detect NVC using backchannel signals. The videos of NVC receiver participants are tracked and features extracted using *geometric-a* features, as described in Section 4.2.3. A classifier is trained on the sender NVC labels and the receiver facial feature data. Testing is performed on a person independent basis, with only clear NVC clips used in the test E_c^{clear} (see Section 4.4) based on annotation ratings of the NVC sender.

The experimental arrangement is almost identical to that presented in the previous chapter. The target classes and annotation data is described in Chapter 3. Feature extraction is performed as discussed in Section 4.2. Classification is performed as described in Section 4.3. Eight fold cross validation is used on clear examples of NVC (Section 4.4) and performance is evaluated using AUC, as described in Section 4.5. The only difference is the shape features used for training and testing are taken from the other person in the dyad (i.e. the person conversing with the subject that was shown to the annotators).

Table 5.4: AUC ROC performance using backchannel features (clip level testing, person independent testing, average score of categories and average standard deviation from Tables A.9 to A.12 are shown).

Test	Multi-person		Person independent	
	SVM	Adaboost	SVM	Adaboost
affine	0.50 ± 0.04	0.53 ± 0.05	0.49 ± 0.05	0.50 ± 0.04
deform-cubica	0.51 ± 0.01	0.52 ± 0.05	0.50 ± 0.00	0.49 ± 0.06
deform-fastica	0.51 ± 0.01	0.52 ± 0.05	0.50 ± 0.00	0.49 ± 0.06
deform-pca	0.54 ± 0.04	0.62 ± 0.07	0.50 ± 0.00	0.52 ± 0.04
geometric-h	0.65 ± 0.05	0.66 ± 0.07	0.54 ± 0.04	0.59 ± 0.07
geometric-a	0.67 ± 0.03	0.68 ± 0.09	0.59 ± 0.08	0.59 ± 0.09
lbp	0.54 ± 0.05	0.58 ± 0.05	0.49 ± 0.07	0.48 ± 0.10
lm	0.55 ± 0.03	0.56 ± 0.06	0.48 ± 0.02	0.50 ± 0.05

5.3.1 Results and Discussion

The performance of automatic classification based on backchannel NVC signals is shown in Table 5.4. If the performance for forward channel NVC classification (Table 4.3) is compared to backchannel performance, it can be seen that the backchannel performance is equal or lower in every case. This is unsurprising, because backchannel information is only a behavioural response to forward channel communication. The results suggest that, for some methods of feature extraction , the classifier is consistently performing at above chance level of 0.5. However, additional tests are required to establish if this result is statistically significant. If the performance for each of the feature extraction methods is compared, it can be seen that *geometric-a* features are again the most effective. The majority of other feature sets have performance levels at or near chance level. Person independent testing is more challenging than multi-person testing in all cases. These observations have the same general pattern as the forward channel performance results.

One difference between backchannel and forward channel is that SVM classification has equal to or worse performance when compared to Adaboost. This may be due to over

fitting of the SVM, which might be addressed by parameter tuning.

5.4 Conclusion

This chapter has identified coupled behaviour in facial expression and described a study of backchannel features to classify forward channel NVC. This was based on previous research that found many human behaviours are inter-personally coordinated in two person conversations. Coupled facial deformations were identified for both corresponding regions of the face, as well as non-corresponding regions. The corresponding areas are situated in the mouth area or were thought to relate to head pitch. Classification was also possible using backchannel features, although the performance is significantly lower if compared to forward channel features.

This area of research is still at an early stage. There are many potential improvements and extensions to the work presented in this chapter. For instance, coupled behaviours were only examined in conversation pairs, which may lead to personal differences in behavioural patterns. It would be possible to modify the method to look for behavioural patterns that consistently occur across a larger group of conversation participants. Also, only instantaneous face deformations were considered. If an effective way of encoding face deformation or gesture was found, it may be possible to check for responses to stimuli that occur after a time delay, in a similar fashion to Richardson et al. [262]. Pearson's correlation coefficient is only sensitive to linear correlations between variables, but it is quite possible that non-linear behavioural patters may exist. Only facial deformation is considered in the features and it would be interesting to expand the behavioural encoding to verbal communication, as well as arm position and body pose. This would enable a much broader range of social phenomena to be studied.

There are also possible extensions to classification based on backchannel features. It should be possible to combine backchannel and forward channel features, either by feature or decision fusion, to improve performance. Also, many of the limitations of the previous chapter apply here: NVC is not temporally modelled, and the social context and cultural background of the participants and annotators could be better controlled.

To address the cultural differences of the annotators, the next section describes the collection of a culturally specific annotation set for the TwoTalk corpus.

There is no truth. There is only perception.

Gustave Flaubert

6

Multiple Culture Annotation

Humans often disagree as to the meaning that is expressed in a video clip of NVC. As discussed in Section 3.6, the statistical mean of ratings is used to form a single label for each video clip. If there is significant inter-annotator disagreement in the data, taking the mean is not necessarily valid because the resulting label does not correspond to any specific human or group of human annotators. One of the factors that affects NVC signal perception by humans is the cultural background of the observer (see Section 2.1). As the goal of automatic recognition of NVC is to produce labels similar to those generated by humans, modelling the specific cultural perception of NVC signals will lead to improvement over training a system in one culture and then applying it in a different culture.

For the purposes of automatic recognition, all previous research has either gathered annotation data from a specific culture or merged annotations from multiple cultures.

This is the first study using subjective annotation for machine learning so that cultures were kept as distinct annotation sets.

The main contributions of this chapter are:

- Annotation data, based on crowdsourcing, that is suitable basis for automatic recognition of human behaviour.
- A quantitative analysis of cross cultural annotation data. The data is analysed in terms of quantifying the possible existence and extent of differences between cultural specific perceptions.

Research relating to cultural differences and crowdsourcing are discussed in the next section. Section 6.2 describes the Internet based collection of annotation data. This data is filtered to remove uncooperative workers; the filtering method is described in Section 6.3. Section 6.4 describes an analysis of the filtered data.

6.1 Related Research

The field of cross cultural study of emotion was founded by Darwin in 1872 [71], in which he was struck by the remarkable similarity in emotion expression styles across cultures. Darwin also pioneered the use of photographs in the study of emotion, and the assessment of emotion by judgement based studies [98]. The key role of culture in non-verbal communication was stressed by Hall [133], with particular attention to cultural influences of which the expresser is not consciously aware. Ekman modelled the process of differences in emotion expression by proposing culturally specific display rules [93], but claimed that emotion was largely culturally universal, and this universality was particularly evident in a subset of emotions called "basic emotions" [96]. There is less evidence for the universality of non-verbal communication but emotion can be a form of communication [120], and that component of NVC may be partly universal.

Cultural differences in expression of emotion and NVC have been studied and observed [201]. Extreme cases of cultural difference in perception include different head motions for agreement and disagreement [159], and differences in obscene gestures [166]. Gaze

aversion direction while thinking has cultural differences [206]. Backchannel signals are more frequent in Japanese culture than US culture [331]. These differences may be significant for an automatic system if it is trained for one culture and deployed in a different culture. Various cross cultural studies have collected observer judgements and each time have found cultural differences. This has included judgements of pictograms [56] and emotional avatars [199, 168]. Marsh et al. found that observers could distinguish a person's nationality by their style of emotional expression [197]. Japanese observers were found to be more sensitive to context than Westerners in emotion recognition [200]. Some universality was found for non-verbal vocalisation of the common Ekman 6 basic emotions but cultural differences were found in non basic emotions [274]. Even the process of perception has been seen to have cultural differences: gaze during emotion recognition were found to have different patterns [149], which possibly leads to some cultures having difficulty distinguishing certain emotions. All these findings suggest that differences in expression and perception exist but also there is a great deal of commonality between cultures.

Elfenbein and Ambady [104, 103] argued that members of a cultural group are more accurate at recognising emotions in their group, when compared to an out-of-group observer. In response to this paper, Matsumoto argued that for the group hypothesis to be properly tested, the sample sizes needed to be balanced and the stimuli had to be equivalent [201] and this had not been done. The potential for culturally specific concepts can make NVC and emotion words difficult to translate or transfer to new cultures (see Section 6.2.1). If an in-group advantage exists for NVC, cultures that differ from the expresser's culture may annotate data differently. This may range from a reduction in inter-annotator agreement to significant misinterpretations of the NVC meaning.

Various approaches exist to detect and remove outliers in questionnaire responses [347]. The method proposed in this chapter differs in that it considers each annotator response set as something to be accepted or rejected as an entirety. This property makes it more suitable to a crowdsourcing application.

6.1.1 Crowdsourcing of Annotation Data

Crowdsourcing refers to the participation of a loosely defined group of people who work towards a common goal, often using the Internet to coordinate the work. Some scientific projects have used crowdsourcing and involving many workers from across the world, including Cooper et al. [60], for "foldit" protein folding and Riddick et al. for "Galaxy Zoo" [254] which generated enough interest to not need payment to incentivise participation. Tarasov et al. [302] proposed using crowdsourcing to gather emotional label annotation data. There are few or no previous works that use crowdsourcing for annotation of NVC behaviour, for the purposes of automatic recognition. The next section describes how crowdsourcing was applied to annotation of NVC samples.

6.2 Data Collection Method

The annotation task involves viewing a series of short views from the corpus (as described in Section 3.2) and providing responses to the questions on NVC signals (described in Section 3.3). Computer based annotation is commonly used for annotation because the annotation task can be precisely defined by the experimenter, and because of the convenience of collection and analysis of the results. Computer based annotation may also be conducted remotely, which is cost effective and convenient but provides less control on how the annotators complete the task. Differences in the equipment, the presence or absence of audio and a different physical environment may affect how an annotator perceives an NVC signal. These factors could not be controlled using remote computer based annotation and may result in an increase of inter-annotator differences.

The annotation data was collected from multiple cultures. Crowdsourced workers may be located anywhere in the world and they use any standards compliant web browser to complete tasks which have been defined by a work supplier. There is usually little screening of workers and no qualification or requirements beyond being able to access the Internet. A typical screen shot of the web page presented to the annotator is shown in Figure 3.5. Although it is quite possible to independently establish a website that

enables annotation and payments, there are several web-based services that manage the website infrastructure for a fee. Crowdflower's web service[1] was used to manage the annotation work. This service provides a high level interface to other crowdsourcing services, such as Amazon Mechanical Turk[2] and Samasource[3]. Each service has an associated worker pool with a distinct worker demographic. However, crowdsourcing has several issues that need consideration if high quality annotation is required.

As previously mentioned, the order of the questions presented to the annotators was randomised to reduce the effect of question order. It was impossible to insert a demographic survey before the users undertook the main task. This was unfortunate, because demographic information may be useful in accounting for inter-annotator differences. However, the IP address of each annotator was available, which was used to assign a rough geographic location of each Internet user. Also, Samasource specialises in refugee populations and an annotators physical location can differ from their cultural background. In particular, a significant response was received from annotators located in Kenya, but these people were likely to be Sudanese refugees. Also, an IP address is not a perfectly reliable method to localise users, given the existence of Internet proxy services. Despite these factors, an IP address based location was considered sufficient for our needs. In the case of refugees, it is not necessary to know the exact culture of origin, as long as the IP address location corresponds to multiple distinct cultural groups.

Volunteers may be sampled from a sub-group within each culture and this effect is called "sampling bias". The availability of skills and equipment varies in each population, therefore if computer use is required, sample bias is likely to occur. The extent of this bias is unknown. It is possible that there is no unified NVC perception consensus in a single culture, if any intra-culture factor has a large effect on NVC perception. It may therefore be undesirable to sample from an entire country's population, if the annotations within a group are not in agreement. However, using a sub-group of a culture is satisfactory, as long as the sample is not assumed to represent the culture as a whole. If the applications of this technology are to be adopted by a sub-group of a culture, such

[1] http://crowdflower.com/

[2] https://www.mturk.com/

[3] http://samasource.org/

as those with easy access to computers, this bias may be beneficial in creating a system that operates well with potential users. In conventional surveys, various techniques can be used to improve annotator agreement, such as training, annotators previewing the corpus, panel-based ratings, repeat ratings of a clip by the same annotator. These were not used because of the technical limitations of the present crowd sourcing tools, for which work unit presentation is pseudo-random. Annotator normalisation was not used because the annotation data is sparse and perception labels of individual annotators is not required for this study, but this technique may be adapted to this situation in future work.

The survey data described in Section 3.5 was collected using unpaid volunteers. This chapter relies on paid workers to provide annotation data. Given that it was possible for an annotator to provide random data in return for a monetary reward, there can be quality control issues which need to be managed. The process for filtering valid work from random annotation is described in Section 6.3. The workers that accept to complete the task as instructed are designated as "trusted" workers, and annotators that respond with poor data or random data are designated as "untrusted". Crowdflower provides its own semi-automatic tool for identifying trusted workers called "gold", presumably from the term "gold standard". This gold tool was used to reduce expenditure on poor annotation data. However, this tool did not affect the final annotation data and both trusted and untrusted data was retained and filtered using the method described below.

An alternative approach to identification of trusted workers is to introduce additional validation questions with a known answer. Unfortunately, there are several practical problems with this approach in a crowdsourcing environment. Additional questions increase the amount of work required of the annotators. Due to the current technical limitations of the annotation service, each annotation task or "work unit" must have the same number of responses for every question. The basic work unit requires the four NVC signal ratings of the questionnaire. To add one more question would increase the workload by 25%. Also, humans that intend to cheat will be able to identify any validation questions and simply answer them correctly, while randomly answering the NVC questions. This possibility of circumventing the validation questions makes

their usefulness questionable. The annotation data would still need to be analysed and filtered to provide confidence that the annotation data is valid. The next section discusses the issue of language in the context of NVC signal perception ratings.

6.2.1 Translation of Cross Cultural Instruments

The design of the questionnaire was previously discussed in Section 3.3 but previously only a single culture and a single language have been considered. Different languages are spoken throughout the world with English being the first language for approximately 380 million people [3]. Second language speakers and learners of English outnumber first language users but the majority of people speak languages other than English. The questionnaire was applied to cultures that have significant differences, including language usage. The translation of survey instruments is used in many scientific fields to avoid the instrument being understood differently outside of its original culture. According to Geisinger [122], an instrument needs to be translated and the fact that the translated instrument measures the same constructs as the original version needs to be verified. A translation may lead to the instrument measuring something other than what was intended [253], and this makes comparison between cultures problematic. Translation of survey instruments is strongly recommended in medical surveys, using rigorous methods [124, 29, 214]. Medical surveys contain questions that refer to concepts that are labels for observable phenomena that exist separately from the observer. However, the NVC questionnaire is quite different in that the concepts do not exist except as subjective interpretations. The questionnaire asks for an annotator's interpretation of NVC signals which depends on their cultural background, etc. The term "NVC" implies that these communication signals do not directly correspond to word based concepts. De Mendoza [73] argues emotion labels used to express subjective interpretations are not "defined by necessary and sufficient features" but rather "probabilistic concepts with an internal structure with better and worse examples of the category and fuzzy boundaries". In this view, a particular instance of an emotion might be perceived as mostly agreeing, somewhat blaming, slightly surprised or any other combination of fuzzy overlapping labels. The same can be said of NVC signal perception, in that the labels used in the questionnaire are interrelated, fuzzy, partly

overlapping and not comprehensive.

Given that an instrument is to be used in multiple cultures, the concepts referred to in the instrument should be consistently understood. There are two approaches that were considered: translating a questionnaire to cultures of interest or to present a single questionnaire in multiple cultures. These approaches will now be discussed in more detail.

If the concepts referred to exist only in the mind of an observer, these concepts might be "cultural concepts" that do not necessarily exist in another language or culture. Because no one-to-one mapping exists between different cultural concepts, the translation of instruments and cross culture recognition are problematic [103]. There are many examples of emotional concepts that do not have direct translations (*vergüenza* from Spanish to English, *shimcheong* from Korean to English [73]). Applying this to NVC, a particular communication action may be perceived in Spain as a particular combination of cultural concepts but in the United Kingdom, a different combination of cultural concepts. With this culturally specific mapping between concept labels and experiences, a perfect one-to-one mapping will be impossible. Therefore, cultural differences could be caused by imperfect translation leading to different cultural concepts used by the annotators or the way emotions are mapped onto the cultural concepts, with no easy way to distinguish these two effects.

The other approach is to present the same questionnaire in multiple cultures. The words in the questionnaire would be interpreted by the annotators in relation to their own cultural concepts and this can change the basis by which an annotator perceives and evaluates NVC signals.

Given both approaches have problems, the latter option of having a single survey in one language presented to multiple cultures was used. This decision was based on the fact that, given the crowdsourcing tools available at the time, a specific group of annotators could not be targeted by a tailored questionnaire, although this was planned in a future version of the tools. It may be possible to create a questionnaire using word clouds that would lend themselves better to probabilistic translation (e.g. English word cloud to Spanish word cloud). Given all these considerations, the transfer of everyday concepts

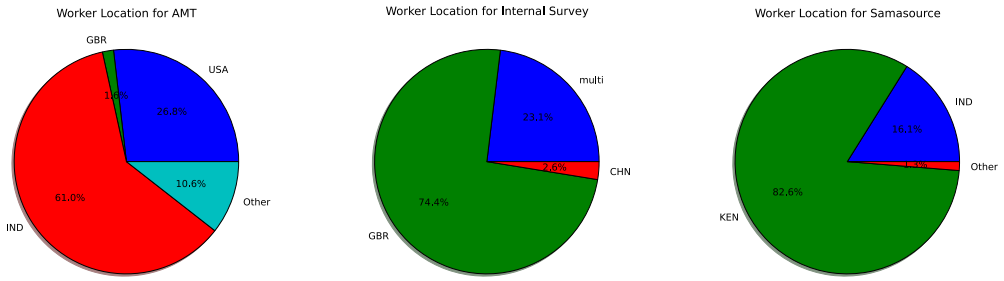

Figure 6.1: Demographics of the worker pools. AMT is Amazon Mechanical Turk. Each country has been abbreviated to its ISO 3166 code. GBR Great Britain, KEN Kenya, IND India, USA United States of America.

from one culture to another is complex and a subject worthy of further study. The next section takes the collected annotation data and considers how untrusted annotators can be identified and removed from the data set.

6.3 Filtering Untrusted Workers

The annotation was performed by 711 participants who provided 79130 individual ratings. These were distributed across 527 video clips and 4 NVC categories, with a total of 2108 questions. However, a significant proportion of the data was random, due to a significant proportion of uncooperative annotators. Three distinct worker pools were used. The internal worker pool is primarily the data previously collected and discussed in Section 3.5, with the inclusion of further annotation by additional volunteers. In all, annotators from 33 countries participated in the annotation task. As can be seen in Figure 6.1, the demographics of each pool differ quite dramatically. Because the annotation data for a single annotator is usually sparse (meaning they are not required to complete every question in the survey), a significant amount of annotation is required for each culture before there is sufficient data to determine stable consensus vote ratings for every question. A considerable amount of annotation data was received from GBR (Great Britain), KEN (Kenya) and IND (India) and these cultures are used throughout the remainder of this thesis. The choice of these cultures is based on the need for

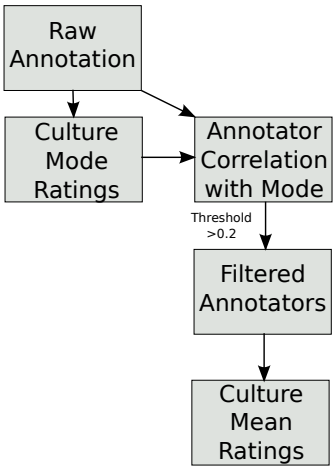

Figure 6.2: Flow chart of filtering method. The statistical mode rating provides a robust standard by which annotators are assigned a trusted or untrusted score.

distinct cultural groups of any origin, for which annotation data is available. The data from the USA was discarded, along with other cultures with a low level of participation, because the annotators of these cultures did not complete enough questions to form a complete set of results. With better targeting of annotation resources, it would be possible to greatly increase the number of distinct cultures.

The workers were divided into trusted and untrusted groups depending on their co-operation with the task. If the data is not filtered, the resultant labels are extremely noisy and are not appropriate for machine learning. The next section describes the filter method used to ensure the labels are of sufficient quality.

6.3.1 Filtering Method

Annotators were divided into two groups: trusted and untrusted. To achieve this, each worker's data $\mathbf{V}_{i,c} \in \mathbb{R}^{d\times 4}$ is compared to a robust rating standard $\mathbf{U} \in \mathbb{R}^{o\times 4\times 3}$. If a worker correlates with the robust standard to a sufficient degree, they are assigned to the trusted group and if not, the untrusted group (see Figure 6.2). The ratings for a single annotator $i \in \{0...d\}$ is found ($m \in \{1...o\}$):

$$\mathbf{V}_{i,c} = \{\mathbf{r}_j : (\mathbf{r}_j, \mathbf{s}_j) \in \mathbf{N}_{c,m}, \mathbf{s}_j = i\} \tag{6.1}$$

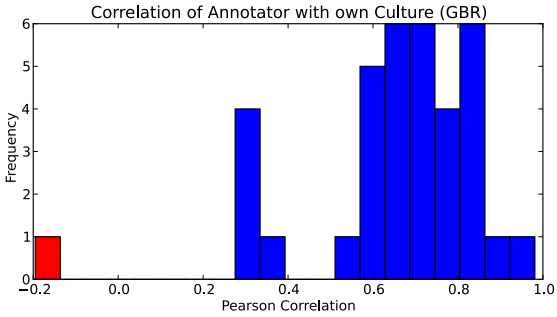

Figure 6.3: Histogram of annotator correlation ρ for GBR when compared to the robust ratings $\mathbf{U}$. Blue bars indicate annotators above the trusted threshold w, and red bars indicate workers below. The majority of annotators in the GBR group have relatively good agreement with the robust rating.

Annotator i is a member of culture $\mathbf{c}_i \in C = \{GBR, KEN, IND\}$. In this chapter, the raw annotation data $\mathbf{N}$ refers to the cross cultural annotation data. The robust rating standard is based on the statistical mode μ for clip m ratings, in NVC category $c \in N$, $x \in C$:

$$\mathbf{U}_{m,c,x} = \mu(\{\mathbf{r}_i : (\mathbf{r}_i, \mathbf{s}_i) \in \mathbf{N}_{c,m}, \mathbf{c}_i = x\}) \tag{6.2}$$

The statistical mode is used because it is relatively robust to uniformly distributed noise. Note that the raw annotation $\mathbf{N}$ is usually considered as a interval-scale, dimensional variable but here it is considered as a quantised variable. This is possible because the annotators' responses used a Likert scale with a finite number of options. Histograms of annotator correlation from the three cultures of interest are shown in Figures 6.3, 6.4 and 6.5. As can be seen, there are cultural differences in the proportion of annotators assigned to the trusted and untrusted groups W. IND had a significant proportion of untrusted workers, while GBR and KEN were largely assigned to the trusted group. These perceptual cultural differences are probably more associated with the crowdsourcing worker pools. The vast majority of untrusted workers were in the Amazon Mechanical Turk pool, while annotators from the Samasource and Internal pools were largely trusted. A specific annotator i provided ratings for a set of clips

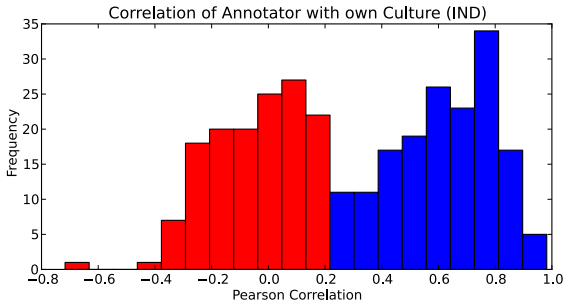

Figure 6.4: Histogram of annotator correlation ρ for IND when compared to the robust ratings $\mathbf{U}$. Blue bars indicate annotators above the trusted threshold w, and red bars indicate workers below. A significant number of workers in the IND group were assigned to the untrusted group.

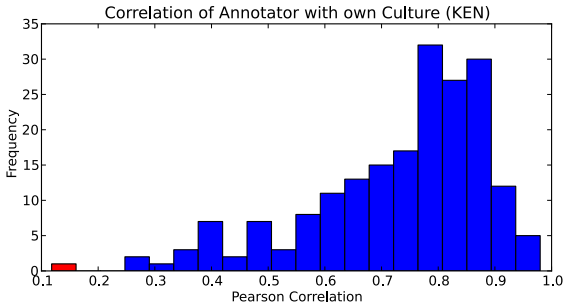

Figure 6.5: Histogram of annotator correlation ρ for KEN when compared to the robust ratings $\mathbf{U}$. Blue bars indicate annotators above the trusted threshold w, and red bars indicate workers below. The majority of annotators in the KEN group have relatively good agreement with the robust rating.

F_i. Annotators only usually rate a subset of the corpus. The matrix of ratings by annotator i is designated as $\widehat{\mathbf{V}}_{i,c} \in \mathbb{R}^{F_i \times 4}$, and the robust ratings that correspond with this clip subset is designated $\widehat{\mathbf{U}}_{i,c,x} \in \mathbb{R}^{F_i \times 4 \times 3}$:

$$\widehat{\mathbf{U}}_{i,c,x} = \{\mathbf{U}_{m,c,x} : m \in F_i\} \tag{6.3}$$

$$\widehat{\mathbf{V}}_{i,c} = \{\mathbf{V}_{m,c} : m \in F_i\} \tag{6.4}$$

The annotator ratings are then compared to the robust standard using Pearson correlation coefficient ρ, $i \in \{0...d\}$. In this case, a 2D matrix, element wise Pearson correlation function ρ' to find correlation $\mathbf{a}_i \in \mathbb{R}$ is used for convenience:

$$\mathbf{a}_i = \rho'(\widehat{\mathbf{V}}, \widehat{\mathbf{U}}_{\mathbf{c}_i}) \tag{6.5}$$

Annotators that have a correlation above threshold $w \in \mathbb{R}$ are assigned to the trusted set of workers:

$$W_{trusted} = (W_{all,i} : \mathbf{a}_i \geq w) \tag{6.6}$$

$$W_{untrusted} = (W_{all,i} : \mathbf{a}_i < w) \tag{6.7}$$

A threshold value of $w = 0.2$ was determined based on the minimum correlation of cooperative annotators from the GBR culture (Figure 6.3). The final filtered annotation $\mathbf{I}$ is computed by taking the statistical mean of trusted annotators. Trusted annotator ratings for a single question form a distribution. The mode only considers the peak of this distribution but disregards the other samples. However, using the mean is more sensitive to the overall views of the annotators. The mean is not a robust measure, which is why it is applied after the data is filtered:

$$\mathbf{I}_{\theta}^{m,c} = \overline{\{\mathbf{r}_i : (\mathbf{r}_i, \mathbf{s}_i) \in \mathbf{N}_c^m, i \in W_{trusted}\}} \tag{6.8}$$

Table 6.1: Number of annotators and votes for cultures included in the unfiltered and filtered **I** data set.

Annotator	Num. Annotators			Num. Ratings		
Country	Total	Trusted	Untrusted	Total	Trusted	Untrusted
India	304	167 (55%)	137	37147	22754	14393
Kenya	196	195 (99%)	1	15452	15420	32
GBR	36	26 (72%)	10	8211	8167	44
Total	536	388 (72%)	358 (28%)	60810	46341	14469

where $\theta \in C$. The quantity of annotations in the unfiltered **N** and filtered **I** annotation sets are shown in Table 6.1. The 28% of users that were assigned to the untrusted group accounted for 23% of the unfiltered data. This implies the trusted annotators answered a greater number of questions on average. After filtering, a significant amount of data remains available for use in training and evaluating machine learning techniques. Each annotator was assigned to a single culture by Internet Protocol (IP) address. This information allows a culturally specific robust NVC perception rating ($\mathbf{U}_{GBR}$,$\mathbf{U}_{IND}$,$\mathbf{U}_{KEN}$), which are used to filter annotators to form a final, culture specific, filtered annotation sets $\mathbf{I}^{m,c}_{GBR}$, $\mathbf{I}^{m,c}_{KEN}$ and $\mathbf{I}^{m,c}_{IND}$.

Untrusted annotators were expected to provide uniformly random responses. Filtering the annotators to remove these responses was expected to result in a reduction in variance in the filtered data **I** when compared to the unfiltered data **N**. There is a possibility that a minority subset of self consistent annotators being selected and this would not represent an overall consensus. However, Table 6.1 shows that the majority of annotators are retained in the trusted set, which shows that this is not the case.

6.4 Analysis of Annotations

The previous section has described how the annotation data was collected from Internet workers and filtering was applied to obtain a high quality set of NVC annotations. This annotation data describes the NVC content for either *agree*, *thinking*, *question* or *understand* as observed from a particular culture (GBR, KEN or IND). However

Table 6.2: Inter-culture correlation of filtered consensus data for different culture pairs. $\rho_{IND,KEN} = \rho(\mathbf{I}^{m,c}_{IND}, \mathbf{I}^{m,c}_{KEN})$. Individual annotators are not compared in this calculation but rather the overall cultural consensus.

	India	Kenya	GBR
India	1	0.56	0.55
Kenya		1	0.64
GBR			1

it is well known that NVC perception is dependent on cultural rules and the specifics of cultural factors on NVC perception is not well understood. This section provides a quantitative analysis as to the nature and extent of the cultural differences in the annotation data.

As can be seen in Table 6.2, the correlation of pairs of cultures is at an intermediate level. A correlation value of less than one (corresponding to perfect correlation) was expected, because previous studies of emotion have observed cultural differences (see Section 6.1). A Pearson correlation above zero (corresponding to no correlation) implies NVC perception in different cultures is not totally independent and a degree of commonality exists. This confirms our suspicion that NVC perception is similar to emotion perception in that cultural differences exist, but there remains a significant commonality between cultures. Another point that is illustrated by Table 6.2 is that cultures have varying degrees of similarity to other cultures for perception of NVC. The KEN-GBR correlation is higher than either IND-GBR or IND-KEN (0.64 vs. 0.56 or 0.55). This suggests that IND is the most distinctive in NVC perception, and KEN-GBR are relatively similar. It might be expected that other cultures are much more or much less similar than the three cultures studied. This work was restricted to annotators with access to computer and Internet resources, but a broader examination of cultural differences would be interesting.

Pearson's correlation coefficient ignores scaling differences when comparing data. If cultures differed by only scaling differences, this would be ignored by the correlation measure. This is a desirable property, because scaling differences between cultures are relatively trivial to understand and model. However, less than perfect correlation scores

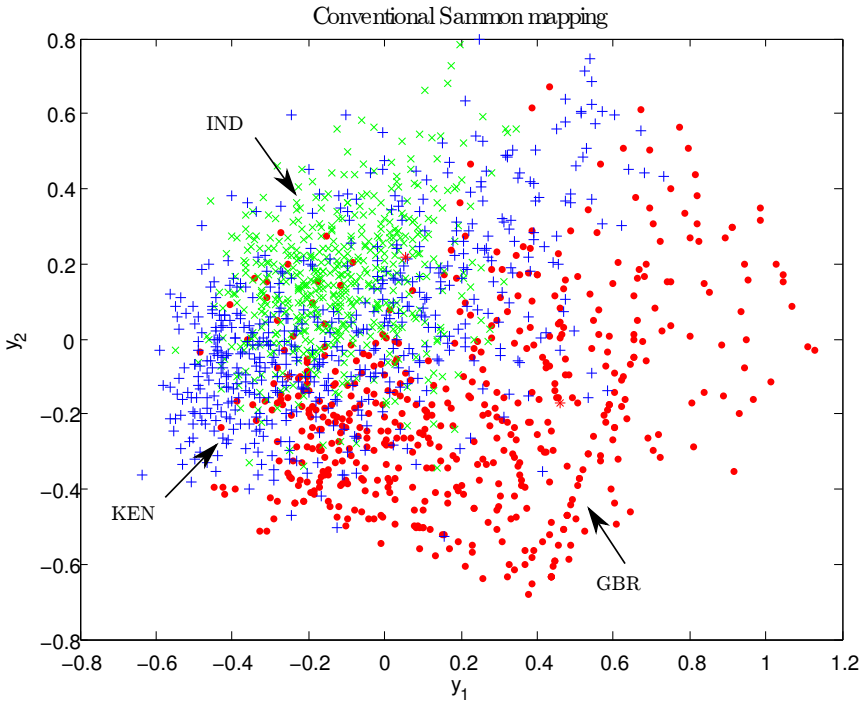

Figure 6.6: Sammon mapping of the filtered annotation responses. Each point represents the mean rating of a single clip within a single culture. The four NVC categories are concatenated into a 4 component vector to enable distance pairs to be computed. GBR: red (.), IND: green (x), KEN: blue (+).

indicate that annotation differences exist and they are not merely scaling differences.

A way of visualising cultural differences in annotation of NVC perception is the Sammon mapping [272]. This technique is a dimensionality reduction tool which maps high dimensional samples to a lower dimensional space while preserving inter-sample distances. For culture θ, the filtered culture annotation $\mathbf{I}_{\theta}^{m,c}$ is a 527 by 4 matrix. The three filtered cultures (GBR, IND, KEN) are combined into a $3 \times 527 = 1581$ by 4 matrix. These points are transformed into a 2D space using Sammon mapping and plotted. This allows us to compare the distribution of annotation responses for different cultures. The result of this procedure is shown in Figure 6.6. This produces a complex distribution with some regions that are generally exclusive to a single culture, areas that are densely occupied by two overlapping cultures and one central zone where all cultures have a high density. The area of three culture overlap corresponds

Table 6.3: Correlation of the mean annotator correlation within various cultures with their respective culture filtered annotation **I** or the global Mean (taken as the combined India, Kenya and UK ratings). Annotators correlate better on average with their own culture consensus than the global consensus **K**.

	India	Kenya	UK
Own Culture Mean	0.67	0.78	0.77
Global Mean	0.64	0.74	0.68

to annotation responses that appear in all cultures, the most common of which likely corresponds to "no NVC is being expressed". Areas that are dominated by a single culture are interesting because they imply that for a subset of clips, a culture has provided a specific rating for the 4 NVC categories that does not appear in the other cultures. This seems to be most apparent in the GBR annotations which has a significant proportion of points that are far from any other culture's points. This supports the idea that cultural differences in NVC are not trivial to explain or model, and any mapping between culturally specific annotation sets may be non-linear.

Because various filtering methods and mean ratings are used to arrive at the filtered annotations **I**, the extent the filtered annotations are representative of the individual annotator responses should be verified. Also, the validity of disregarding culture and the combining ratings to produce a global mean **K** can be considered. These comparisons are shown in Table 6.3 and show an annotator's correlation to their own culture is relatively good. The lower correlation of IND may indicate a problem with data quality or perhaps there is greater perception individuality for this cultural group. However, if cultural difference is ignored as in the case of global mean annotation **K**, the individuals do not correlate as well as with their own culture consensus, in all three cases. Therefore, there is a divergence between individual annotators and the ratings that are intended to reflect them. Remember that taking a mean of a group of ratings is only valid if the annotators form a relatively self-consistent group. Therefore, ignoring culture produces a global annotation that has less validity. If global annotation data is used to train an automatic system, the labels that would be predicted would not necessarily correspond to any single annotator or any subset of annotators. This would

not be compatible with the objective of producing an automatic system that would predict NVC as a human would perceive NVC.

6.4.1 Inter-annotator Agreement

The annotation consensus score $\mathbf{I}_{\theta}^{m,c}$ is formed by taking the mean of trusted annotator ratings. This approach is only valid of sufficient annotators are in agreement with the label. Given that the annotator ratings are dimensional, the level of inter-annotator agreement can be determined by taking the standard deviation of ratings for each culture. A low standard deviation implies a high level of inter-annotator agreement. Figures 6.7 and 6.8 show the cumulative number of clips at various levels of agreement. As can be seen in these plots, some NVC signals have a higher level of inter-annotator agreement (*question*) and others have a lower level of agreement (such as *thinking*). Different cultures are again seen to have an overall higher level of inter-annotator agreement (GBR) than others (IND). This may be caused by different levels of homogeneity in perception of NVC for each culture.

A standard deviation of 0.289 corresponds to the level of agreement for random ratings (based on uniformly distributed ratings between 0 and 1). There are a significant number of *thinking* NVC examples that have a lower level of agreement than would be expected by change. This suggests that there may be more than one perception mode for *thinking* NVC, for at least some of the video clips in the corpus.

Section 7.6 uses clips with a high level of inter-annotator agreement as the basis for regression. This addresses the issue that annotators sometimes do not agree enough to provide a meaningful label.

6.5 Conclusion

This chapter has described the collection of cross cultural NVC perception data. The data was collected based on paid Internet workers from multiple cultures. Because some workers did not cooperate with the task, the data was filtered to isolate the valid annotation data. The annotations were analysed and found to contain differences that

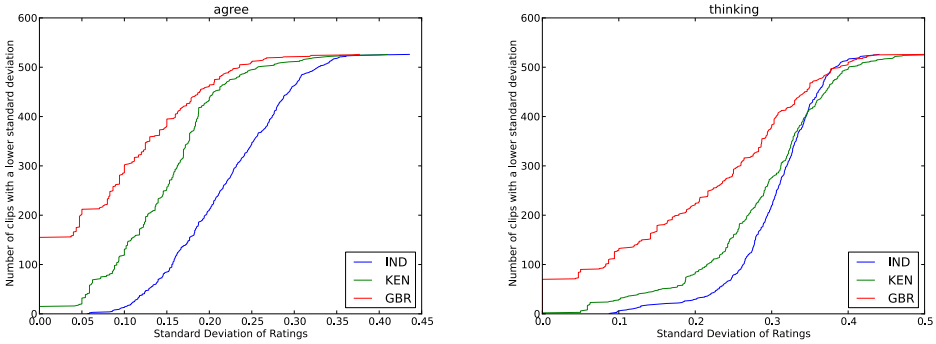

Figure 6.7: Cumulative plot of the number of clips at various annotation standard deviations of annotator ratings. Left plot shows *agree* and the right plot shows *thinking*.

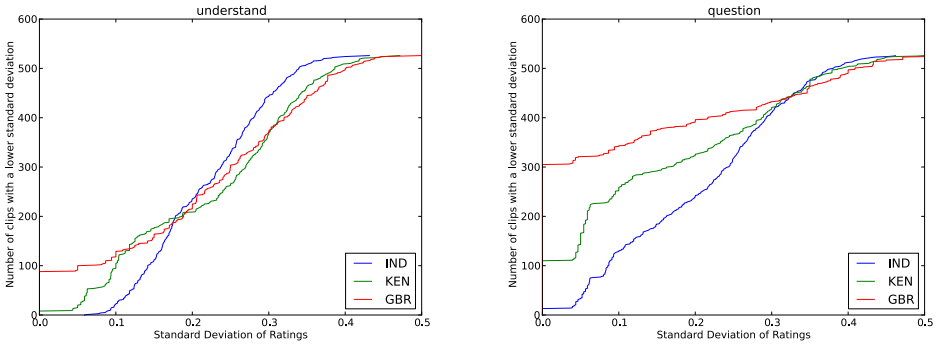

Figure 6.8: Cumulative plot of the number of clips at various annotation standard deviations of annotator ratings. Left plot shows *understand* and the right plot shows *question*.

were not merely linear scaling effects. This video annotation resulted in a new, cross cultural annotation of NVC on natural informal conversations. The annotation data has been publicly released for further use. This helps to address the lack of publicly available NVC annotation data.

Crowdsourcing annotation data may prove to be a significant resource in human behaviour understanding if the quality issues can be addressed. The quality assurance tools that are currently available focus on annotation of discrete, objective labels. These objective labels are appropriate for computer vision tasks such as object recognition where the concepts are well defined. However, discrete label filtering method is not useful in subjective tasks, such as NVC annotation. In this case, inter-annotator disagreements do not necessarily imply that some of the annotators are wrong. Quality is still a significant concern and the filtering method proposed in this chapter is a step towards addressing this. The filtering method does rely on using the annotation mode, which depends on sufficient annotator participation to form a stable result. Also, the annotators are hard assigned to trusted and untrusted groups. It may be more efficient to use soft assignments for annotator trust.

The presence of cultural perception differences is not particularly surprising given a similar effect is often observed in emotion. The significance of the findings in this chapter is the confirmation of NVC annotation differences and a quantitative analysis of the differences. The results suggest that different cultures have varying degrees of similarity. These perceptual cultural similarities might be exploited to reduce the number of culturally specific perception models needed for NVC recognition, possibly by treating actual annotator group perceptions as a mixture of two or more exemplar perception models.

There are significant shortcomings of Internet crowdsourcing for studying perception. The annotators are not screened and the environment in which the annotation is performed is not controlled. Each culture has a different availability of computer resources, different levels of computer literacy, different demographics that will undertake the task and the annotation is performed in different settings. These differences are known to be a factor in emotion perception and are likely to play a role in NVC perception. The

use of demographic questions may allow some of these variables to be controlled. This chapter has used English language questionnaires, but there can be cultural differences in the understanding of language. Given that perception differences in this chapter may be caused by either language perception differences or NVC perception differences, it is hard to distinguish these effects. One possible method to check and possibly quantify language differences in crowdsource annotation is to conduct cross cultural annotation of data that has objective labels, using a single language questionnaire. If the annotators were consistent across cultures, this result would validate the approach of using a single language. Translation of NVC survey instruments replaces the uncertainty of language perception with uncertainty over the validity of the translation of NVC concepts, which by definition do not directly correspond to word based concepts.

The next chapter uses this dimensional, continuous valued annotation data to approach the NVC recognition problem by cross cultural regression.

Behaviour is the mirror in which everyone shows their image.

Johann Wolfgang von Goethe

7

Multiple Culture NVC Regression

The previous chapter has described the collection of NVC signal annotations from three cultures: Great Britain (GBR), Kenya (KEN) and India (IND). This chapter will now apply this data to automatic NVC regression. The three cultures present in the annotation data are treated as three separate ground truth sets. The annotation data is not combined into a global label set because this results in a label set that does not reflect the annotator perceptions (see Table 6.3). This is because culturally specific annotation data better reflects the annotator perception of NVC used to create the data set (see Section 6.4). Also, discrete groups of annotators enables us to train an NVC recognition system with one culture's annotation set and test it using a different culture. This will experimentally show the performance of using training data that is dissimilar to the test culture. Most emotion and NVC studies do not consider cultural differences but it is expected that the training and test data needs to culturally correspond to achieve good performance. This raises a potential problem, if current methods are

deployed in different cultures without appropriately accounting for cultural difference.

The work builds upon the NVC classifier described in Chapter 4, but changes are introduced to address some of the previous chapter's shortcomings. In summary, these improvements are:

- use of dimensional, continuous valued labels, which makes this task a regression problem,
- use all the samples in the corpus, including intermediate intensity NVC and examples of disagreement, rather than just the clear samples,
- use of a slightly different machine learning technique: ν-SVR (pronounced nu-SVR),
- use of annotation data from three distinct cultures,
- the performance is evaluated by Pearson's correlation,
- only a single feature set is used (*geometric-a*),
- clips of various lengths are summarised into a fixed length "digest" vector by taking the feature mean and variance and
- only person independent testing is performed.

In Chapter 4, the problem was simplified to two classes by considering only clear examples of each NVC signal. This greatly reduces the amount of available training data. Every sample in the corpus is used in this chapter, with dimensional, continuous valued labels (see Section 3.3). This increases the difficulty of the task because intermediate strength NVCs can be difficult to distinguish. This has previously been discussed, with respect to gaze and *thinking*, in Section 4.8.

Because NVC recognition is now expressed as a regression problem, different machine learning techniques must be used. Adaboost and standard SVMs are binary classifiers and cannot be directly applied. Other techniques may be applied to regression, such as ν-SVR [278] and this is described in the next section.

Three culturally specific annotation sets are used, spanning four NVC signals (*agree*, *thinking*, *question* and *understand*). For the 3 cultures and 4 NVC categories, $3 \cdot 4 = 12$ separate regression models are constructed and evaluated independently. It would be desirable achieve a good overall performance because an effective method should generalise to many different types of NVC and operate in different cultures.

Feature tracking and feature extraction are used to transform digital frames into a form usable for effective machine learning. The method used is based on computing a statistical features based on the mean and variance of individual feature components. This encodes general activity level and facial expression without encoding the order of events. A statistical feature is a feature that is based on the statistical properties of a set of unordered observations. Statistical features are also known as temporal features [51, 209, 211], as well as ensemble features. Only a single feature extraction approach is used in this chapter (*geometric-a*), to limit the number of experiments to a manageable scale. This technique was selected because it had the highest overall performance and generalised well to unseen subjects in Chapter 4.

The approach involves NVC regression of each label component based on a distinct, culture and NVC specific models. There may be a wide range of possible interpretations for perceiving and interpreting NVC (see Section 2.1) because interpretation is dependent on context. Given the number of potential factors that can affect perception, the number of specialised models may become too large to create or use in a practical application. However, the problem is not as insurmountable as it may at first appear.

- For a specific application of automatic NVC recognition, there are probably a limited range of social and cultural situations which will be of interest.
- There are culturally and socially based similarities between humans, which greatly reduces the scale of the task.
- General models can be used as a basis and adapted to users during a brief, person specific training session. A similar approach is used in automatic speech recognition, in which a general model is adapted to a user's unique style of speaking.

Although it would be a vast undertaking to comprehensively solve the problem in every

situation, NVC recognition based on specific groups of annotators and in a specific application is feasible.

The main contributions of this chapter are:

- A study of recognition based on culturally distinct sets of annotators to better model culturally specific NVC perception. The automatic system uses dimensional, continuous valued NVC labels which provide richer information content than discrete labels.
- An analysis of the effect of training with one culture's annotations and testing on a different culture's annotation data.

The next section discusses the existing studies relevant to this work. Section 7.2 provides an overview of the automatic system. Section 7.3 discusses how statistical feature are extracted and how regression is applied. Results are presented and discussed in Section 7.4.

7.1 Related Research

This section reviews the literature relating to facial analysis regression, temporal encoding by feature extraction and the use of multiple annotator groups. While there are many papers that use classification of emotion, there are relatively few that use regression. Many regression methods have been proposed but this section will focus on methods that have been applied to facial analysis. The most popular regression method is SVR [86] that extended the SVM algorithm, which was originally formulated as a binary classifier [318]. Emotion recognition has occasionally used dimensional, continuous valued labels such as valence, activation and dominance. Grimm et al. [131] used SVR to predict these labels based on audio input. Kanluan et al. [158] also used these labels for multimodal emotion recognition, with the modes combined by decision level fusion. They found that some label components were better recognized by audio and others by visual information. Some methods focus on modelling temporal variations while making dimensional, continuous valued predictions. Nicolaou et al. [223] used the

related output associative RVM technique for regression of valance-activation labels. Wöllmer et al. [333] found that neural networks can exceed the performance of SVR for emotion regression. Both papers conclude that temporal modelling of features is important. Rudovic et al. [269] used a prototypical emotion model to predict labels for different intensity emotions from the BU-4DFE dataset.

The previous paragraph has outlined papers that address emotion. There are few papers that use non-emotion labels based on facial analysis. Lucey et al. [189] used pain as a dimensional, continuous valued label of internal state. Savran et al. [276] used regression for FACS AU regression using a hybrid 2D/3D approach. Rudovic et al. used a Laplacian-regularised Kernel Conditional Ordinal Random Field model for AU intensity regression [268]. Yang et al. [338] and Kim and Pavlovic [162] approached the same problem using the concept of ranking of ordinal labels. Facial expression, pain, and emotion are often expressed without any conscious intention to communicate. Regression has not previously been applied to NVC meaning.

As many authors have stressed, temporal encoding or modelling of human behaviour is critical to good recognition performance. One approach is to encode temporal variations by feature extraction . A simple way to achieve this is to calculate statistics of features in a temporal window. Datcu and Rothkrantz [72] calculated variance of feature components in a sliding window and compared this to classification of individual frames, finding temporal encoding improves performance. Valstar et al. [317] used many forms of statistical measures, including symmetry. They found that measures that encoded maximal speed and displacement of the face, as well as order of co-occurrence were most useful in distinguishing posed from spontaneous brow movements. Petridis and Pantic used standard deviation and quadratic curve fitting of each of the feature components [246] in a temporal window. Grimm et al. [131] and Wöllmer et al. [333] computed statistical measures on audio features. Fang compares video sequences by using the mean of feature components [109]. All these papers found that temporal encoding by feature extraction was an effective approach. However, the best encoding method may be task dependent. If many statistical measures are computed, the feature vector can become excessively large and this may require feature selection to isolate the important components.

Several corpuses have used multiple annotators, as discussed in Section 3.1.2. Typically, the annotation data is combined to form a single set of consensus labels. Hardly any studies consider groups of annotators as distinct and separate, and use this property in a machine learning context. Perhaps the only previous example of this is Reidsma's thesis [257] which uses a subset of self-consistent annotators to train an automatic system that better reflects the perceptions of the annotator subset. However, this subset is not known to correspond to any particular meaningful group. Reidsma and op den Akker [260] considered each annotator separately and trained an independent model for each. Based on a test sample, each model votes to produce a final prediction by decision fusion. However, this approach is not used in this study because of the difficulty in motivating annotators to rate a significant proportion of the corpus which is required to train effective regression models. No existing work uses subjective labels from subsets of annotators that correspond to culturally or socially identifiable groupings, and then applies this as a basis for machine learning.

7.2 Overview

This section provides an overview of the automatic recognition system. The main steps are depicted in Figure 7.1. LP tracking is again used (see Section 4.2.1). The tracking of each frame is then used to create a frame feature $\mathbf{f}_{alg}$ using the *geometric-a* method (see Section 4.2.3). The features are normalised on a per-subject basis, as defined in Equations 4.2 to 4.4. For each annotated clip, these frame features are combined into a clip digest vector $\mathbf{E}$ using simple statistical measures. Each clip has a single digest vector and a matrix of labels $\mathbf{I}$ (as described in Chapter 6). Eight fold cross validation is performed by dividing clips into person independent training and test sets. For each training fold, a ν-SVR model is trained for each of the 12 components of $\mathbf{I}$. The ν-SVR models are then used to predict labels for the unseen samples in the test set. These predicted labels are compared to ground truth and the performance is evaluated by Pearson's correlation (in a similar fashion to [131, 158]).

The details for feature extraction used to form the clip digest vector will be discussed in the next section, as well as some of the properties of ν-SVR.

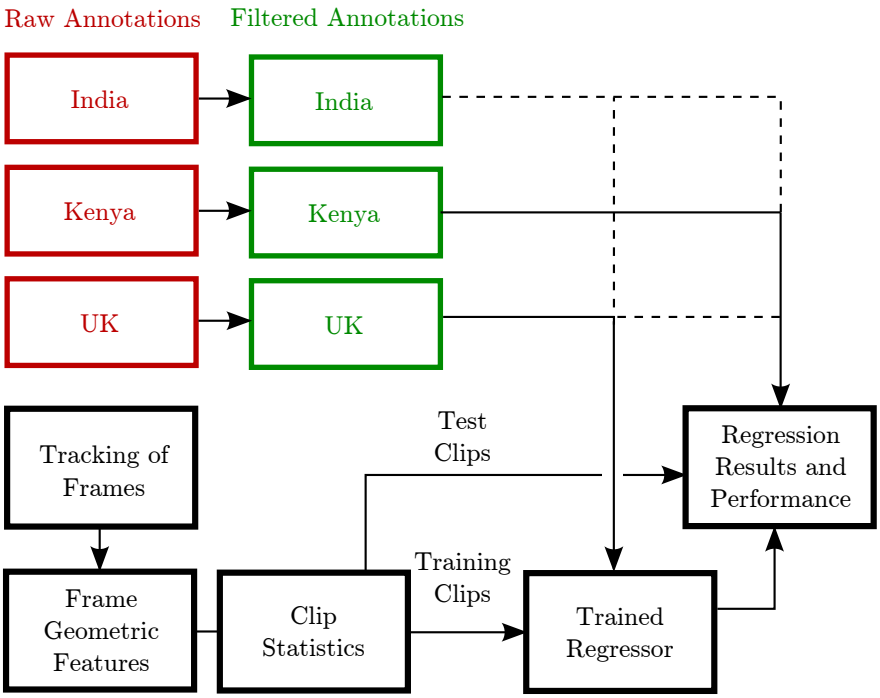

Figure 7.1: Overview of automatic system, showing filtering of annotations followed by training and testing on separate cultures.

7.3 Statistical Feature Extraction and Regression

This section describes the details of the regression system and the differences from the classification approach discussed in Chapter 4. Tracking and feature extraction is used to transform video frames into a form that can be used by standard machine learning techniques. As before, feature point tracking and feature extraction are performed by Linear Predictor trackers and *geometric-a* features, previously introduced in Chapter 4. However these features consider individual frames separately and do not consider how the features vary over time. The clip digest vector, described below, encodes the feature variation over time.

7.3.1 Computing the Clip Digest Vector by Feature Extraction

Video clips in the corpus have variable lengths but an automatic system requires a way to compare the similarity of clips. Clip level and frame level performance measurement was introduced in Section 4.5. However, this approach was not particularly satisfactory

because the frames were considered in isolation to produce a prediction by decision fusion. It would be preferable to consider how features vary across multiple frames. For this reason, the clip digest $\mathbf{E}$ is computed by taking the mean and variance of each feature component. The clip feature matrix $\mathbf{B} \in \mathbb{R}^{r \times s}$, is summarised into a digest vector $\mathbf{E} \in \mathbb{R}^{2 \cdot s}$. This fixed length vector $\mathbf{E}$ may then be used as an input to many standard machine learning tools, including ν-SVR. For feature component i,

$$\mathbf{E}_i^m = \overline{\mathbf{B}_i^m}, i \in \{1...s\} \tag{7.1}$$

$$\mathbf{E}_{i+s}^m = \phi(\mathbf{B}_{i+s}^m), i \in \{1...s\} \tag{7.2}$$

where $\phi(\mathbf{x})$ is the statistical variance of vector $\mathbf{x}$. This is analogous to temporal quadratic features in Section 4.7, but instead of fitting a quadratic curve to a feature component, a Gaussian distribution is fitted to the frame samples. This decision was also based on the motion of eyes during thinking and the visualisation of feature space in Section 4.8. The main finding of that work was that there is little consistency in eye movement except for overall spread and offset from the origin, which can easily be encoded by taking the mean and variance of features. Neither this digest vector, nor the previous approaches of clip or frame level comparisons consider the order of the frames. However, it would be incorrect to say that these features are not temporal, because they do encode feature variation in multiple consecutive frames. Methods that consider the order of the frames have already been discussed elsewhere (e.g. polynomial features and Akakın and Sankur's work on HMMs and CRFs in Section 4.9). Based on these findings, the claim that the frame order is a panacea for all NVC and emotion recognition problems cannot be justified.

7.3.2 Support Vector Regression and Performance Evaluation

ν-SVR [278] is used to perform regression of NVC signals. ν-SVR is similar to SVMs in that it uses weighted kernels centred on sample points to specify a non-linear transform from feature space to a space that is more suitable. In the case of ν-SVR, the distance from the feature space hyperplane is used to perform the regression, rather than just to

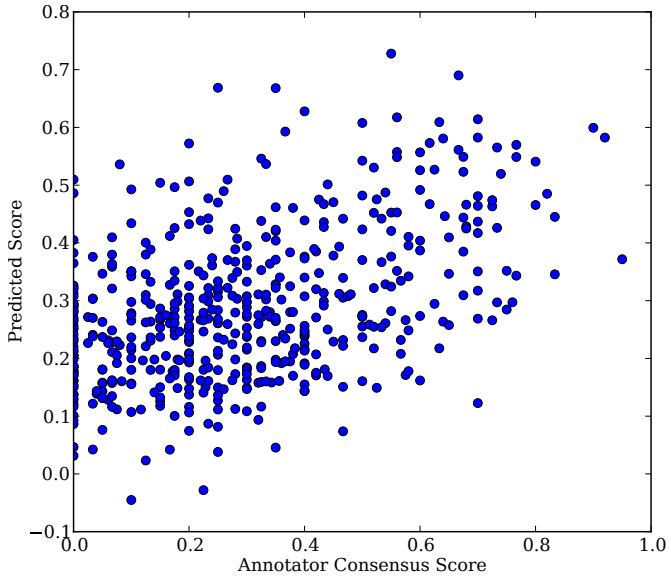

Figure 7.2: Scatter plot of ground truth and automatically predicted *thinking* NVC intensities for the UK culture. Each point corresponds to a single video clip.

separate the space into two regions as done by SVMs. As in the case of classification, an RBF kernel is used.

Because the labels are dimensional, continuous valued variables, the method used to assess performance must be considered. ROC analysis, as used in previous chapters, only applies to two class problems. A performance metric that operates on dimensional, continuous valued variables is required. Pearson's correlation coefficient is used, as previously introduced in Section 3.6. Person independent testing is conducted, rather than multi-person testing, because the person independent case has a broader range of applications than multi-person testing and because it is more challenging.

7.4 Results and Discussion

This section discusses the experimental results of the automatic NVC regression system. The parameters for ν-SVR of $C = 1.0$ and $\nu = 0.5$ where found to be effective.

The predicted labels and ground truth labels are plotted in Figure 7.2 for a single NVC category (*thinking*) and for a single culture *GBR*. A perfect system would have

Table 7.1: Correlation of automatic system for training and testing on a single culture. The error limits are one standard deviation of the individual cross validation correlation results. Figure 7.2 shows the individual ratings for UK thinking NVC. Note that this table uses a different performance metric than in Chapter 4, therefore the performance values cannot be directly compared.

Culture	Agree	Question	Thinking	Understand	Average
India	0.38±0.11	0.20±0.14	0.33±0.13	0.23±0.13	0.29
Kenya	0.43±0.15	0.15±0.19	0.39±0.20	0.43±0.17	0.35
GBR	0.27±0.08	0.16±0.21	0.46±0.20	0.37±0.13	0.32
Average	0.36	0.17	0.39	0.34	

a linear arrangement of sample points. As can be seen, the distribution of points is certainly not linear, but an overall trend can be observed. The Pearson correlation of this distribution is 0.46 (a score of 1 being perfect correlation and 0 indicating no correlation). There are more samples on the left than on the right area of the plot. This is related to the annotated frequency of different intensity NVC signals (see Figure 3.10). The majority of points fall close to where a linear regression line would fall, which implies that most samples have a relatively good NVC prediction.

Figure 7.2 shows only one of the NVC signals and in a single culture. With annotation data from multiple cultures on four NVC categories, the performance of different NVC signal regression models can be compared. As can be seen in Table 7.1, the performance of some NVC signals is low, particularly *question* NVC. The difficulty of *question* recognition was previously seen for classification in Section 4.6. Similarly, *thinking* has the best performance in both regression and classification (see Table 4.4). This is unsurprising because both approaches use similar feature extraction and machine learning techniques. Intense expressions of *question* NVC are relatively rare in this corpus and the regression algorithm may have difficulty learning a general model if insufficient positive samples are provided. In contrast, *thinking* is relatively common and has a distinct visual appearance, making it easier for a visual system to recognise (see Section 4.8).

Each culture has a different average performance. This may be because one or both of:

- the culture's annotators use facial features that are encoded with varying degrees of success by *geometric-a* features.
- the quality of annotation varies across culture. IND culture had the largest proportion of untrusted annotators (see Section 6.3.1) which might indicate a quality problem. It is also possible that different cultures may contain different levels of diversity in NVC perception which may affect the validity of the consensus labels **I**.

The annotation labels are based on a mean consensus score **I** for each question (described in Section 6.3.1). However, individual annotators often differ from these consensus ratings. The human annotator correlation with the culture consensus is shown in Figure 6.3. It is questionable if exceeding these performances would be meaningful, because the labels would not correspond to any specific, observable human perception of NVC. The human correlation therefore provides an upper ceiling on the performance of our automatic system. This could be addressed by having a focused subset of annotators that have a higher inter-annotator agreement.

The performance of NVC signal recognition is at an intermediate level and certainly far from being a reliable prediction. This is due to the extremely challenging nature of the task. The confounding factors include:

- most NVC signals are quite subtle and can be masked by larger face deformation cause by emotion and head pose changes,
- human behaviour in spontaneous conversations is much more variable than posed behaviour (Section 1.2), so the feature space to label partitioning is likely to be complex,
- spontaneous behaviour is hard to track, which increases input noise and
- some NVC signals are based on verbal information which is not considered in this system.

A system with a performance such as this may be useful for some applications where robustness is not a critical requirement. For example, in video retrieval, a list of

Table 7.2: Correlation of performance of the automatic system when training and testing on the same or different cultures.

Test Culture	Train Culture		
	India	Kenya	GBR
India	0.29	0.27	0.24
Kenya	0.27	0.35	0.33
GBR	0.25	0.33	0.32

candidates for clear examples of NVC for a human to make a final section could use a noisy regression system to rank the examples. A wider range of applications could be addressed if better performance can be attained.

Table 7.2 considers the case of training the automatic system on annotations from a single culture (either GBR, KEN or IND) and testing on annotations from a second culture (again either GBR, KEN or IND). The table diagonal, marked with a grey background, contain the performance of the system in the case where the training culture matches the test culture. For these tests, the results from different NVC categories were combined by taking the average performance of the four NVC signal performances. The most important point of this table is the diagonal performance values broadly exceed the performance values that are off diagonal. This shows that training a system in one culture's NVC annotation and then testing on a different culture results in a performance loss. Current practice is to consider automatic NVC and emotion recognition systems in a single culture and a single social situation. These results suggest this will not be an optimal approach if the culture specific model is intended to be used in a wider range of cultures. Although this result might be expected based on cross cultural research, this is the first quantitative measurement of the effect. This table also suggests a possible solution to the problem: to train specialised systems on appropriate cultural perception data to achieve better performance. The implementation of a cross culture NVC recognition system is one of the primary contributions of this thesis.

To illustrate typical predictions that are made by the automatic system, two random video samples will be examined in more depth. The random video clips are shown in Figures 7.3 and 7.4. The first "3dcfiL5Per" clip prediction shows that the predicted

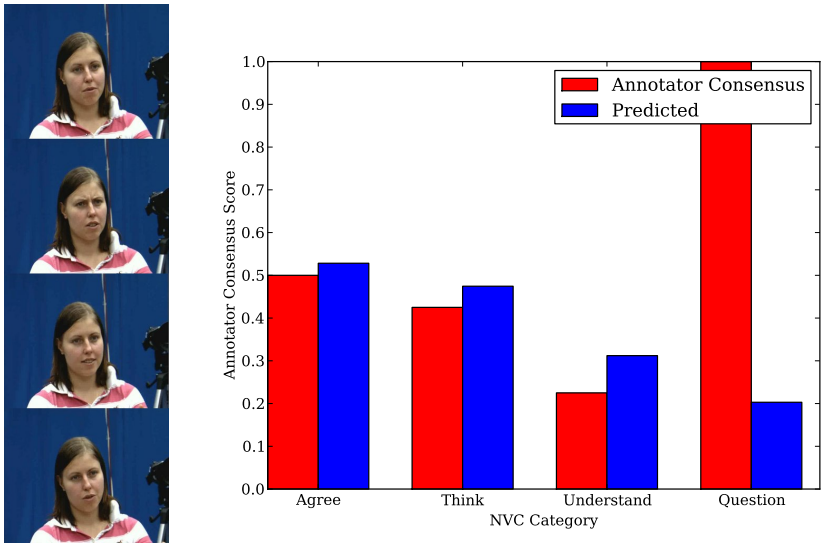

Figure 7.3: Example frames, annotator ratings and predicted scores for corpus clip "3dcfiL5Per", in the GBR culture.

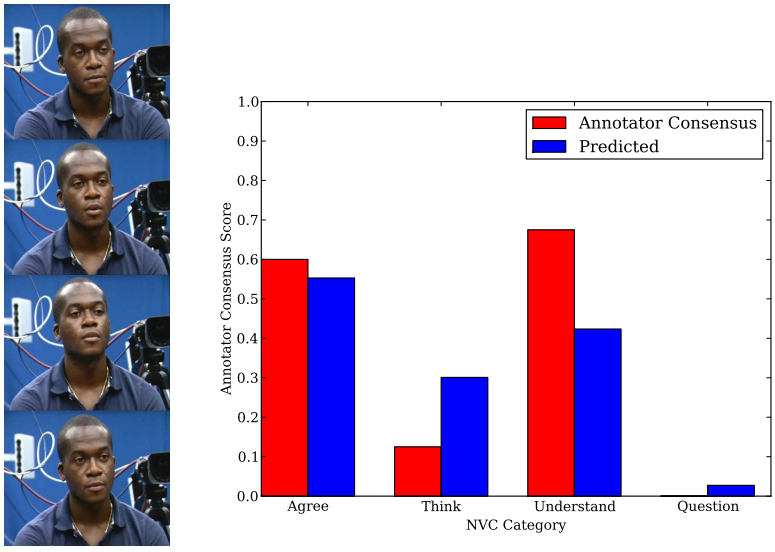

Figure 7.4: Example frames, annotator ratings and predicted scores for corpus clip "GyUrjdl6VT", in the GBR culture

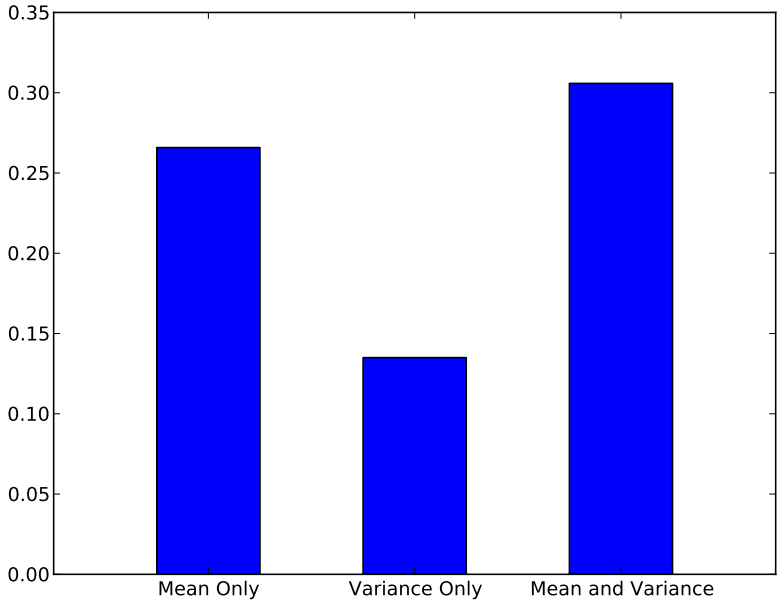

Figure 7.5: The performance of the automatic system using either feature mean statistics or feature variance statistics or the original approach of using both mean and variance statistics. The NVC categories and cultures have been combined by taking the average performance.

(blue) bars are in close agreement with the annotation ratings (red) in three of the NVC categories (*agree*, *thinking* and *understand*). However, the annotators have identified this clip as asking a question, resulting in a high score for *question*. The automatic system, has in this case, failed to provide an appropriate prediction.

The second clip "GyUrjdl6VT" (Figure 7.4) shows that all four predictions are at least approximately correct. The NVC categories *agree* and *question* are almost exactly correct, while the labels for *thinking* and *understand* are in rough agreement. These two figures imply the system is broadly making useful predictions but occasionally has significant errors.

As discussed in Section 7.3.1, feature extraction was performed by taking the clip mean and clip variance of each component of $\mathbf{f}_{alg}$. When either the mean features or variance features were disabled, the performance was reduced (see Figure 7.5). Using

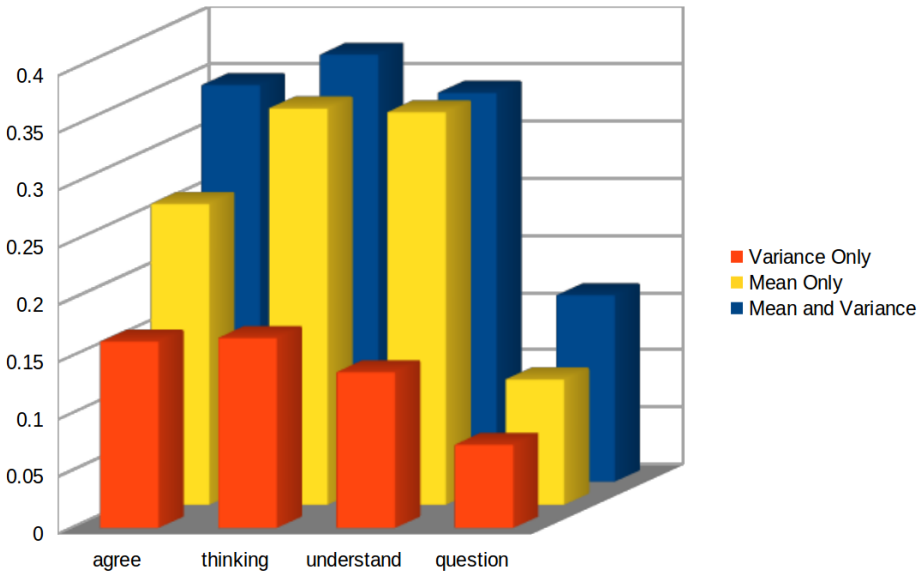

Figure 7.6: The performance of the automatic system using either feature mean statistics or feature variance statistics or the original approach of using both mean and variance statistics. Each NVC category is separately shown.

only mean based features resulted in a higher performance than just using the variance based features. The mean of features encodes facial expression shape, while variance of features encodes how much motion or activity there is in a local area of the face. This suggests that static face expression contains more information than facial activity for some NVC signals. However, there may be more subtle temporal variations that contain relevant NVC information. The best performance is when both of these approaches are combined and implies that facial expression and facial activity contain complementary information. Several papers have noted that temporal information is crucial in effective emotion recognition (see Section 7.1). Based on the results of this chapter, this holds true of NVC as well as for emotion. Perhaps future work can find improved statistical measures that better encode the information to improve system performance.

The role of facial expression and facial activity can be investigated for individual NVC categories. Each NVC involves different gestures and different areas of the face. Figure 7.6 shows the performance of using only mean statistics or only variance statistics for each category. For most NVC categories, both variance and mean statistics provide

the best approach for automatic regression. However, the performance for *understand* is approximately equal if the mean statistics approach is compared to the combined mean and variance approach performance. This implies that variance does not contain any complementary information for *understand* NVC. It is possible that some NVC expressions may be recognized by purely static facial expressions. However, three NVC signals require both facial expression and the temporal variation of expression for optimal performance.

7.5 Applying Digest Vector to NVC Classification

The concept of encoding a video clip into a fixed length digest vector may also be applied to the person dependent classification problem discussed in Chapter 4. The approach described in the preceding section is slightly modified to make it suitable for binary class labels and AUC performance evaluation. As in Chapter 4, only clear examples of NVC are used, which is a less challenging problem than using all samples in the corpus.

Algorithmic features are generated on video frames and normalised as described in Section 4.2.3. A digest vector is then computed for each video clip using the method described in Section 7.3.1. A ν-SVC classifier [278] was trained using two fold cross validation of training and test data. As well as providing a conventional binary label prediction, ν-SVC can provide a prediction of the class membership probability of an unseen sample. This was used in an ROC analysis to determine the AUC performance.

The performance is sensitive to the ν-SVC cost parameter C. The parameter was determined for each fold of the training data independently. Parameter search of cost values of $\{10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}\}$ was performed on the training data on a leave one sample out basis. The cost parameter with the best AUC performance was used to train a final classifier on the entire set of train data.

The performance of the system is shown in Table 7.3. The changes to feature extraction and classification result in a performance increase from 75% to 85.4%. This performance is an improvement on the previous state of the art, which was the overall

Table 7.3: AUC performance, expressed as percentages. Testing is on a multi-person basis. The highlighted row corresponds to the method described in this chapter.

Method	Agree	Question	Thinking	Understand	Average
Akakın and Sankur, 2011 [10]	85.9	78.2	83.8	83.6	82.9
Sheerman-Chase et al. 2009 [287]	70	73	81	80	76
Result from Chapter 4	70	70	83	75	75
Digest Vector, ν-SVC	81.4	89.7	95.8	74.8	85.4

performance of the HMM approach proposed by Akakın and Sankur [10]. There are some NVC signals, such as *agree* and *understand*, for which HMM classification is still more effective than the method proposed here. The overall improvement in performance shows that for some NVC signals, appropriate feature extraction can effectively encode temporal variations. In some cases, this can exceed the performance of some temporal classifier methods.

7.6 Regression on High Inter-annotator Agreement NVC Signals

The annotation consensus score $\mathbf{I}_c^m$ is formed by taking the mean of trusted annotator ratings. This assumes taking the mean of annotator ratings forms a meaningful NVC label. However, the level of inter-annotator agreement, measured by rating variance, is different for each clip (see Section 6.4.1). By removal of clips of low inter-annotator agreement, it may be possible to remove data with meaningless labels from the corpus. However, if the filtering is too aggressive, useful and meaningful data may be discarded and the problem begins to become over-simplified.

Experiments were performed based on the method in Section 7.4 except that only a subset of clips from the corpus was used for training and test. The subset was based on an inter-annotator threshold. Clips with a higher variance than the inter-annotator threshold were excluded. The system was tested with the inter-annotator threshold at various values. The correlation performance of these experiments are shown in Figures 7.7 and 7.8.

As can be seen in these figures, reducing the threshold to focus on clips with higher

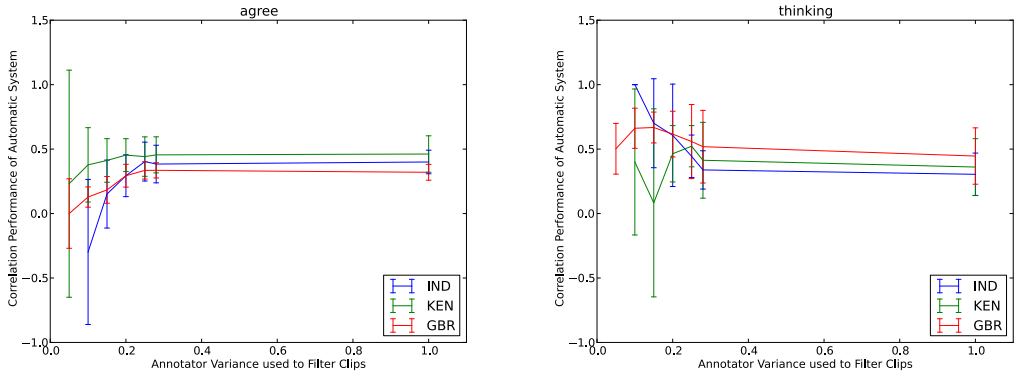

Figure 7.7: Correlation performance of automatic system after clips with a low inter-annotator agreement have been removed. The left plot shows *agree* and the right plot shows *thinking*. Error bars of one standard deviation are shown.

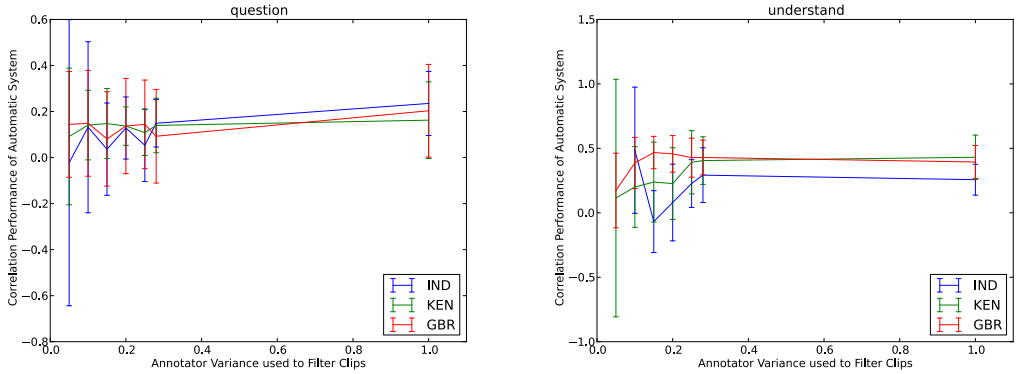

Figure 7.8: Correlation performance of automatic system after clips with a low inter-annotator agreement have been removed. The left plot shows *question* and the right plot shows *understand*. Error bars of one standard deviation are shown.

inter-annotation agreement results in little overall change in performance. The greatest positive effect on performance is for *thinking*, which may be due to the removal of ambiguous, multi-mode perceptions that may be present for this NVC signal (see Section 6.4.1). The effect on *agree* is detrimental, which suggests that the machine learning technique is robust to the label noise that is present, while inter-annotator agreement removes data that is useful for training. The other NVC categories show no change or possibly a sight negative impact on performance.

Filtering the corpus based on inter-annotator agreement does not appear to be an effective way of improving performance in this case. It may be interesting to investigate the characteristics of NVC samples that have higher and lower inter-annotator agreement, as well as the cause of agreement. The appropriateness of agreement metrics as the basis for machine learning could also be explored further [257]. At present, we can say that the performance is relatively independent to the level of inter-annotator agreement filtering and this NVC corpus contains some ambiguous label data which but this is not significantly detrimental to performance.

7.7 Classification of Agreement and Disagreement in the Canal 9 Corpus

The Canal 9 corpus comprises a series of Swiss/French TV studio political debates with a duration of 42 hours. The debates are between 2 to 4 participants and a moderator. The large number of subjects and the use of a single broadcasted view which has been edited from multiple cameras makes tracking more resource intensive. Having more than two participants and the TV cameras may result in increased head pose changes. The corpus has pre-existing annotation data which specifies the shot start/end and the identity of people in the shot. The quality is broadcast TV with interlaced frames. This annotation has been expanded to instances of agreement, disagreement and neutral by Bousmalis et al. [38] for a subset of shots that have a total duration of approximately 10 minutes and featuring 30 subjects. It is unclear if these annotations were intended to encode mental states or communicated meaning.

Table 7.4: Classification balanced accuracy performance for agreement and disagreement for the Canal 9 corpus. The last two rows are the performance of the proposed method. Chance prediction performance is 50% accuracy.

Classifier	Features	Balanced accuracy
SVM	hand/arm/shoulder gestures	50% [38]
SVM	prosody	50% [38]
SVM	hand/arm/shoulder gestures & prosody	52% [38]
HMM	hand/arm/shoulder gestures	43% [38]
HMM	prosody	51% [38]
HMM	hand/arm/shoulder gestures & prosody	52% [38]
HCRF	hand/arm/shoulder gestures	51% [38]
HCRF	prosody	56% [38]
HCRF	hand/arm/shoulder gestures & prosody	64% [38]
NuSVC	facial *geometric-h*	50%
NuSVC	facial *geometric-a*	51%

The problem is framed as a two class problem and evaluated in a five fold, "leave one debate out" cross validation by Bousmalis et al. [38]. They approached this problem by audio prosody feature extraction and manual hand/arm/shoulder gesture annotation, followed by training a temporal model classifier (HMMs or HCRFs) to provide automatic predictions. One static classifier was also used (SVM). Their approach is similar to this thesis in that they do not use verbal information. Interestingly, they provide performance data for gestures, prosody and both combined. The performance metric used was balanced accuracy.

The method in this chapter is adapted to the Canal 9 data by using NuSVC[278] classification, which is closely related to the regression method used earlier. Faces were tracked using an extension of LP tracking described in Sheerman-Chase et al. [288] and *geometric-h*/*geometric-a* geometric features are extracted (see Sections 4.2.2 and 4.2.3). The features are normalised on a per-subject basis, as defined in Equations 4.2 to 4.4. Multiple frames are combined by taking the feature component mean and variance as described in Section 7.3.1. NuSVR is used to provide a predicted class label.

The performance are shown in Table 7.4. Just using the visual modality, including the method proposed in this chapter, results in chance level performance. Bousmalis et

al. achieve the best results using prosody and arm gestures with HCRF. This is likely to be the only result which is significantly above change, although this needs to be confirmed by more experiments. These results are surprising given the above chance classification and regression results using the TwoTalk corpus. There are a number of possible reasons for this result:

- Visual features do not provide enough relevant information for humans or automatic methods to predict agreement and disagreement. (Although seeming not the case for TwoTalk)
- The facial area does contain relevant information, but the increase head pose changes and lower resolution of the face (typically 150 by 250 pixels in interlaced video) make tracking and extracting this information difficult.
- If relevant facial information is extracted, a non-dynamic model may not be suitable for classification. (Although it was suitable for TwoTalk.)
- The Canal 9 social context does not rely on facial behaviour for agreement and disagreement compared to the TwoTalk social context.

Judging by visual inspection, the tracking of the Canal 9 data seems to be relatively good. Either poor tracker of the different social context seems the most likely explanation for the poor performance. This raises the possibility that different social contexts require different approaches and even possibly different modalities to achieve effective behaviour or NVC recognition. This may be an interesting area of future work.

7.8 Classification of Mental States in the Mind Reading Corpus

Mind Reading Emotions Library is a commercial dataset that was originally developed for assisting those on the autism spectrum. The database comprises of 412 “concepts” or classes of mental states. Each concept has 6 silent videos of actors performing the mental state as well as 6 audio recordings. The database was published in 2004 and

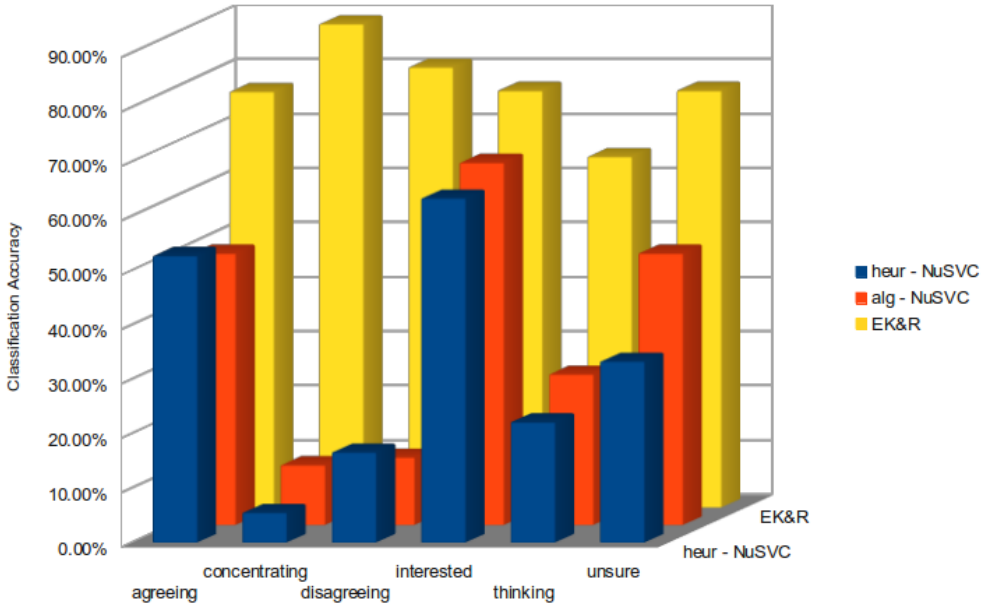

Figure 7.9: Classification accuracy for Mind Reading mental states using various methods. Blue bars correspond to *geometric-h* heuristic features classified by NuSVC in one-against-all fashion, orange bars correspond to *geometric-a* algorithmic features classified by NuSVC in one-against-all fashion and yellow bars correspond to el Kaliouby and Robinson [99]. Chance prediction performance is 17%.

Table 7.5: Confusion matrix and classification accuracy of Mind Reading emotional states from el Kaliouby and Robinson [99], Figure 7. Chance prediction performance is 17%. Rows correspond to the true label, columns to predictions.

mental state	agreeing	concentrating	disagreeing	interested	thinking	unsure	accuracy
agreeing	**26**	4	0	1	0	3	76.5%
concentrating	1	**16**	0	0	0	1	88.9%
disagreeing	1	1	**17**	0	0	2	81.0%
interested	2	2	0	**23**	0	3	76.7%
thinking	1	4	0	3	**20**	3	64.5%
unsure	2	3	1	0	1	**23**	76.7%
mean							**77.4**%

Table 7.6: Confusion matrix and classification accuracy of *geometric-a* algorithmic features classified by NuSVC in one-against-all fashion. Chance prediction performance is 17%. Rows correspond to the true label, columns to predictions.

mental state	agreeing	concentrating	disagreeing	interested	thinking	unsure	accuracy
agreeing	**18**	1	6	3	6	2	50.0%
concentrating	2	**2**	1	5	3	5	11.1%
disagreeing	9	0	**3**	1	8	3	12.5%
interested	2	2	0	**20**	3	3	66.7%
thinking	6	3	4	1	**10**	10	27.8%
unsure	2	1	3	3	8	**15**	50.0%
mean							**36.3**%

the video encoding is quite dated, with low quality compression and a small resolution of the face: typically 100 by 150 pixels. el Kaliouby and Robinson [99] grouped some of these Mind Reading concepts into an arrangement of 6 classes (see Figure 3.2 in [102]). The mental states they studied were *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. This subset has 164 individual clips with a total duration of 19 minutes. These were used to train an automatic system based on a hybrid tracking and appearance features and a multi-scale temporal DBN classifier. Ten sequences were discarded because the face tracker did not correctly initialise the head position for the first frame of the video clip. The particular clips excluded was not specified, making a replication of these experiments difficult. Performance evaluation was conducted on a "leave one clip out" basis and reported as classification accuracy. Their results are reproduced in Table 7.5.

Table 7.7: Confusion matrix and classification accuracy of *geometric-h* heuristic features classified by NuSVC in one-against-all fashion. Chance prediction performance is 17%. Rows correspond to the true label, columns to predictions.

mental state	agreeing	concentrating	disagreeing	interested	thinking	unsure	accuracy
agreeing	**19**	1	6	3	4	3	52.8%
concentrating	1	**1**	1	4	10	1	5.6%
disagreeing	7	0	**4**	2	8	3	16.7%
interested	4	2	1	**19**	0	4	63.3%
thinking	8	5	5	1	**8**	9	22.2%
unsure	2	0	2	4	12	**10**	32.3%
mean							**32.3**%

Although the labels in this corpus are superficially similar to the TwoTalk corpus used in this work, the labels of Mind Reading correspond to mental states and not to NVC (see section 2). However, it is possible to adapt the system proposed earlier in this chapter to the Mind Reading 6 class problem. As before, features are tracked and *geometric-h*/*geometric-a* geometric features are extracted (see Sections 4.2.2 and 4.2.3). The features are normalised on a per-subject basis, as defined in Equations 4.2 to 4.4. Multiple frames are combined by taking the feature component mean and variance as described in Section 7.3.1. NuSVR in a one-vs-all arrangement is used to provide a predicted class label.

The confusion matrices of the proposed method are shown in Tables 7.6 and Table 7.7. A class specific accuracy summary of these results and el Kaliouby and Robinson are shown in Figure 7.9. As can be seen, the performance of the proposed method is significantly lower than that of el Kaliouby and Robinson. In particular, the mental state classes of *concentrating* and *disagreeing* are probably no better than chance predictions. *thinking* is much better than chance, although further experiments are needed to establish the significance of most of these findings. *unsure* and *agreeing* are of intermediate performance but still far below el Kaliouby and Robinson's result. Only the *interested* mental state is nearing el Kaliouby and Robinson's performance. As can be seen in the confusion matrices, *unsure*, *thinking* and *concentrating* are often mistaken for one another. This is hardly surprising, since each of these mental states are very similar in outward behaviour and function. For the proposed method, the *agreeing* accuracy

exceeds the *thinking* performance, while for NVC the *thinking* was recognized more consistently than *agree* (see Figure 7.2). These effects many be cause by:

- Acted behaviour is more intense and more consistent than spontaneous behaviour. This may be exploited by the temporal model used by el Kaliouby and Robinson to achieve a better result.
- NVC and mental states are different concepts. The proposed system was developed for NVC and el Kaliouby and Robinson's system was developed to address the Mind Reading corpus.
- There is less video data available for each class and more subjects in the Mind Reading corpus, which may affect which method is suitable. Specifically, the proposed method uses subject specific normalisation which can require a significant amount of data to reach a stable and robust model of facial behaviour.
- el Kaliouby and Robinson discarded video clips that were not tracked, while we consider all videos, which is a harder problem.
- The feature extraction technique used by each method is different, which may encode relevant information that is missed by other feature extraction approaches. Although the *geometric-h* heuristic features are based on el Kaliouby and Robinson, it is not the same. They use appearance features that focus on mouth opening and teeth visible, which may be useful in distinguishing some of the classes.

The performance of the proposed system may be improved by feature selection or by the use of temporal model based classification. This is discussed in Section 8.4.

7.9 Conclusion

This chapter describes an automatic system that is trained on culturally specific annotation data. This enables the automatic system to better model the cultural differences in NVC perception. The predictions are dimensional, continuous valued labels, which provide richer information than using discrete NVC class labels. All clips are used in

this chapter, including examples of intermediate intensity NVC and are tested in a person independent fashion. Simple geometric features are used, based on distances between pairs of trackers. Temporal information is encoded by taking simple statistical measures of feature components.

Human interpretation can only be performed in a specific context. It is difficult to determine if annotation differences are caused by perception differences, or cultural differences in the use of annotation tools. There are a number of areas in which the current work could be improved. If natural conversations were recorded with NVC encoders located in different cultures, effects of perceiving one's own culture vs. perceiving a different culture could be studied. Also, behaviour differences could be investigated, but it may be difficult to ensure each social situation is equivalent across cultures to enable comparison. Better control of the annotators and encoders would also reduce the possibility that non-cultural effects being the dominant cause in differences of perception. Having the annotators work in a controlled environment and given the same instructions may perhaps improve consistency.

The general approach of specialised modelling of a culture's perception can be scaled to other cultures as long as training data is available. If every possible culture, social situation and personality requires a different model, the high resource requirements to produce these models makes this impractical. However, there are commonalities across social situations and cultures. Therefore, this commonality may be used to adapt an existing model to approximate a new culture or situation without retraining the system from scratch. This process may be similar to adaptation of a speech recognition system to a personalised accent.

Given the evidence that personality and gender affect emotion perception, it is possible these are factors in the perception of NVC. Also, annotation data may be gathered without this information being available. It may still be possible to train a tailored system by unsupervised clustering of annotators based on their responses. This would make NVC label predictions better reflect actual human responses but the personality type or gender of the annotators would be unknown.

The features used are based on an exhaustive extraction of distances between pairs

of trackers. This can lead to the inclusion of irrelevant or redundant features, which can cause problems for machine learning algorithms. This can be addressed by feature selection and is discussed in the next chapter.

Simplicity is the ultimate sophistication.

Leonardo da Vinci

8

Feature Selection for NVC Regression

The features used in the two previous chapters were extracted based on distances between pairs of tracker positions. These features encoded local deformations of the face, with each feature component corresponding to a specific, local area. Expression of emotions or NVC signals only involve part of the face but the features also encode areas of the face that are not involved in a specific NVC expression. Therefore, the feature data contains irrelevant information. Redundant information is also present in the algorithmic features, with tracker motions in a local area of the face often being closely coupled. Training based on features that are largely irrelevant leads to over-fitting, in which the NVC model is based on noise or incorrect features, rather than modelling the true underlying pattern. This results in poor generalisation and poor performance. This can be addressed by various means, including feature extraction that is focused on the task at hand, or by feature selection.

Feature selection is the process of determining the relevance of features for a Machine Learning (ML) task. Using a smaller subset of relevant features reduces the computational complexity of ML. As the relevance of each feature is computed, this can give an insight to how NVC is expressed. Feature selection provides a way to find associations between NVC signals, relevant feature components and the associated areas of the face. Feature selection is a broad discipline with many approaches. Features are either assigned a relevance weight or a subset of relevant features is found. This chapter uses an existing wrapper based approach which determines which subset of features is relevant. The selected feature subset can then be applied to NVC recognition and the performance change can be evaluated.

An NVC signal specific feature selection is computed for each culture. The performance contribution of each feature component is also evaluated. This results in a set of feature relevance weights for each NVC signal. The feature weights can be visualised to show the involvement of facial areas in the expression of NVC in an intuitive manner. This is based on segmenting a face using Voronoi tessellation around the position of trackers. Voronoi tessellation segments an image into cells based around seed positions; each point in the space is assigned to a cell based on the nearest seed position. This visualisation can either be used to check if the relevant facial areas conform to our expectation, or provide an indication as to the areas used by the automatic system. This in turn may provide clues as to human NVC perception, although facial areas used in human perception of emotion may differ from an automatic approach.

The main contributions of this chapter are:

- A method of feature selection to select subset of features that are relevant for NVC recognition.
- A method of visualisation of the relevant features that is easy and intuitive to interpret.

The next section provides an overview of relevant existing research. Section 8.2 discusses feature selection for the purposes of improving performance.

8.1 Related Research

There are many approaches to feature selection, which vary in performance, computational cost and restrictions on the type of input data. There are three general types of feature selection (see Saeys et al. for a review [270]):

- filter methods, that consider feature components either individually or in groups, but do not consider the ML method used to make the final prediction. This scales to large sets of features. These methods do not consider effect of the ML method and may ignore the effects of combining groups of features.
- wrapper methods, that consider sets of features in conjunction with an ML method. They do consider groups of features in context with ML but they tend to be computationally intensive, they are classifier specific and are prone to over fitting.
- embedded methods, which are combined with ML methods to select features. This makes the feature selection methodology and results specific to the ML method used.

Existing papers that use feature selection for facial analysis will now be discussed. The use of an embedded feature selection, such as a boosting classifier, can be used to weight a set of features based on relevance. This feature subset can then be used by a second, more sophisticated classifier. This approach was used by Valstar [317] to select shape features by Gentleboost, and Petridis and Pantic [246] used Adaboost to select relevant audio and visual features. However, performing feature selection in this way, assumes there is similarity in the optimal set of features for both methods, which might not be the case. Yang et al. [340] propose a feature selection method based on rough set theory on audio visual features. This avoids discretisation of feature values, as required by Adaboost, which may result in a loss of information. Filter based feature selection appears to have been largely avoided, probably because the number of feature components in the original feature vector is relatively low (usually thousands of feature components at most), and the interaction between features is often significant for emotion and NVC recognition.

Wrapper based methods include randomised feature selection approaches such as simulated annealing and genetic approaches, but these have not been popular in facial analysis. Deterministic wrapper based approaches have been applied to emotion recognition: Grimm [131] used SFS to isolate relevant audio features. This method begins with an empty set and incrementally adds features that produce the greatest performance increase, in a greedy fashion. An alternative, called Sequential Backward Elimination (SBE), is to start with a full set of features and incrementally eliminate features that result in the best performance [163]. The SBE approach was used by el Kaliouby and Robinson [99] to find the most relevant geometric features.

There are several existing papers that identify which features have been selected for emotion or NVC recognition, but it is less common to attempt to visualise which features have been selected. If features are shown, they are often visualised individually (e.g. [340]), which can make comprehension of the overall distribution difficult. In experimental psychology, gaze patterns in perception have been visualised [149]. Busso and Narayanan visualised areas of the face that were correlated with prosodic and vocal tract features for different emotions [47]. It would be beneficial to have a method that is similar to these approaches for visualisation of feature selection results in a way that can be intuitively understood.

8.2 SBE Feature Selection

The approach used is a greedy SBE of the features [163]. This section describes the method in detail and the resulting performance impact. Backward searching was thought to be preferable to forward searching because the interaction of features can be found and exploited. Forward search, particularly in the first few iterations, adds features without the benefit of other complimentary features. In contrast, a backward search allows irrelevant features to be eliminated while retaining features that contain complementary information. Also, it is possible to accelerate the backward search by removing more than one feature at each iteration, which reduces the computational cost. Some feature component subsets are more effective than others. Unfortunately, SBE may not find a globally optimal feature subset. If all possible feature subsets are

considered as a space, SBE performs a gradient descent to minimise error.

8.2.1 Method

Feature selection occurs within a person independent, cross validation framework. There are eight folds in cross validation, resulting in eight different partitionings of seen and unseen data sets. Feature selection is applied to the seen data of a specific cross validation fold, to determine a relevant feature subset. SVR is then applied to the feature subset to produce a model suitable for prediction.

Algorithm 1 Features are selected by a greedy backward elimination, beginning with a feature set that contains all possible features. The feature set at each stage is stored for later use. The algorithm is expressed as Python 2.7 code. The function *EvaluateRemainingFeatures* is defined in Algorithm 2.

```
def PerformFeatureSelection(data, labels):

    #The current mask begins with every feature enabled
    currentMask = np.ones((data.shape[1]), dtype=np.bool)

    #For storing the intermediate steps during feature selection
    allMasks = []

    #While more than one feature remains in the current mask
    while currentMask.sum() > 1:

        #Evaluate which features to remove
        toRemove = EvaluateRemainingFeatures(currentMask, data, labels)

        #Update the current mask and remove features
        for featToRemove in toRemove:
            currentMask[featToRemove[1]] = False

        #Store mask for later analysis
        allMasks.append(currentMask)

    return allMasks
```

The procedure for SBE is shown in Algorithm 1. The search begins with a current set $\boldsymbol{\alpha} = \{1...s\}$ which is initialised to include all possible feature components. The

components to be removed from $\boldsymbol{\alpha}$ at each iteration is then determined. The current set $\boldsymbol{\alpha}$ is then updated and the process continues until the current set $\boldsymbol{\alpha}$ is empty. For the large number of components, it is too time consuming to remove components at a rate of 1 per iteration. To accelerate the process, multiple feature components are removed nearer the start of the SBE process. As the number of components in the current set approaches zero, the rate of feature elimination returns to the standard 1 feature component per iteration. This produces a significant speed increase, but risks the removal of non-optimal components and this may result in a sub-optimal final feature set. The number of feature components removed from the current feature set at each iteration is denoted η. This depends on the number of feature components ω in the current set $\boldsymbol{\alpha}$ as follows:

$$\eta = \begin{cases} 200 : \omega > 1000 \\ 100 : \omega > 400, \omega \leq 1000 \\ 1 : \omega \leq 400 \end{cases} \tag{8.1}$$

These thresholds were based on an intuitive expectation that only a small subset of features are required for accurate recognition.

To find an appropriate subset of features for removal from the current feature set, the contribution of each feature component needs to be assessed. An overview of this process is shown in Algorithm 2. Each feature component in the current feature set $\boldsymbol{\alpha}$ is selected as the test component and the performance impact of the removal of the component is evaluated. The features are then prioritised, with the feature components resulting in the lowest performance preferred for removal. This process becomes progressively faster as the current feature set becomes smaller.

To test a specific feature component, regression models are trained and tested to assess the performance of the remaining feature components as shown in Algorithm 3. A temporary set $\boldsymbol{\beta}$ is created which creates a copy of the current set $\boldsymbol{\alpha}$ except for the removal of the test component i:

Algorithm 2 Each feature in the current feature set is tested. The features that cause the best performance are retained and the features that worsen performance are preferred for removal. The function *TestPerf* is defined in Algorithm 3.

```
def EvaluateRemainingFeatures(mask, data, labels):

    #Determine how many components to remove, based on the mask
    numToRemove = CalcNumToRemove(mask.sum())

    #Initalise an empty list, based on the number of components in the mask
    scores = [None for count in range(mask.sum())]

    #For each remaining component in the mask
    for count, featNum in enumerate(np.where(mask == True)[0]):

        #Create a test mask with a single component disabled
        testMask = np.copy(currentMask)
        testMask[featNum] = False

        #Evaluate the perfomance of the test mask and store the score
        scores[count] = (TestPerf(data[:, testMask], labels), featNum)

    #Return a list of features to remove
    scores.sort()
    scores.reverse()
    return scores[:numToRemove]
```

Algorithm 3 Testing the performance of a test set of feature components. The regression can use any suitable method, but in this study ν-SVR is used.

```
def TestPerf(maskedData, labels):

    #Prepare cross validation sets
    kf = cross_validation.KFold(len(maskedData), numFolds)

    #Create an empty list for predictions
    allPredictions = []
    allLabels = []

    for train, test in kf: #For each cross validation fold

        #Fit a regression model on the training data
        regressor.fit(maskedData[train,:], labels[train])

        #Store the predictions based on the test data
        allPredictions.extend(regressor.predict(maskedData[test,:]))

        #Store the test labels in this fold
        allLabels.extend(labels[test])

    #Calculate the overall correlation of predictions and test labels
    return np.corrcoef(allPredictions, allLabels)[0,1]
```

$$\boldsymbol{\beta} = \{j : j \in \boldsymbol{\alpha}, j \neq i\} \tag{8.2}$$

The feature data is split into cross validation folds. These "feature selection" folds are distinct from the "system" cross validation folds discussed earlier, so that each fold contains data from multiple human subjects.

This produces a series of sets $\{\boldsymbol{\alpha}^s..\boldsymbol{\alpha}^1\}$ that correspond to each stage in the progressive removal of features. Each set contains a different number of feature components. The performance of the feature set on the unseen data can then be determined. The expectation is for an increase in performance as poor features are removed. As the SBE process is nearing termination, some features that are critical to NVC regression are removed and the performance sharply declines. The performance of the feature subset at each stage is evaluated and retained for later analysis.

Because this process results in multiple sets which are used to create multiple NVC models, it is not obvious which feature set to use and how many feature components are optimal. Simply selecting the peak performance when evaluating feature sets on unseen data violates the separation of seen and unseen data. For simplicity, this section uses the feature set $\boldsymbol{\alpha}^{unseen}$ having the peak performance for unseen test data to determine the number of feature components. It is likely that different NVC signals require a specific set of geometric features to be effective. Therefore, feature selection is computed for a specific NVC category and using a specific culture's annotation data. The processing of test set $\boldsymbol{\beta}$ has been parallelized in this implementation, resulting in a speed increase. The next section uses the set $\boldsymbol{\alpha}^{seen}$ having the highest performance on seen training data only.

8.2.2 Results

A typical plot of performance against the number of feature components in the subset is shown in Figure 8.1. As expected, the performance of predicting unseen test data increases as features are removed until performance suffers a sharp decline. The far left starting point of the lower curve corresponds to the performance of the system

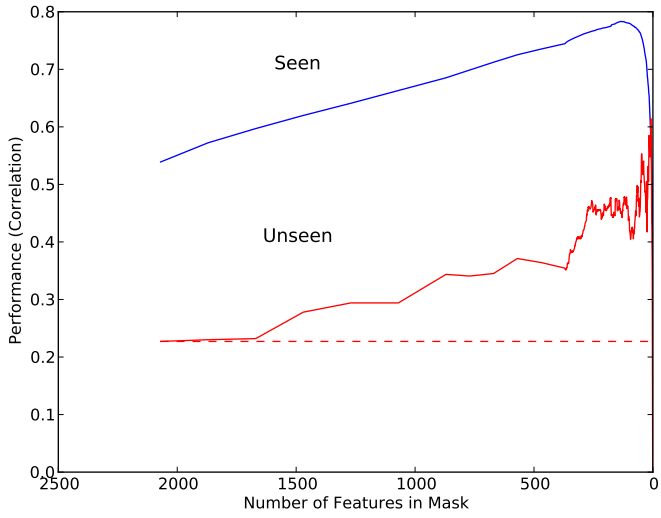

Figure 8.1: The performance of the system progressively improves as backward feature selection eliminates poor features. The upper line shows the seen data, which is used in the feature selection algorithm. The lower line shows the performance of the unseen data. The unseen data corresponds to subject 3008.

discussed in the previous chapter (i.e. without feature selection). In this example, feature selection results in a significant increase in performance. A feature set containing between 10 to 275 features encompasses the highest performance, with the peak performance requiring only 10 features. This feature set is smaller than the terse $\mathbf{f}_{geometric-h}$ features (Section 4.2.2). However, it is unlikely that this feature set will be effective for other non-*thinking* NVC regression. This can be a disadvantage because the SBE method is NVC category specific and a great deal of computation is required to retrain the system for a different NVC signal.

The feature selection curves are relatively linear until approximately 400 features remain. This corresponds to the threshold in Equation 8.1. The change in the curve behaviour at this point suggests that a different set of thresholds might result in a higher peak, although this was not investigated.

This pattern is repeated for most other NVC categories and in different cultures. While almost every test fold subject benefits from the feature selection process, not all system cross validation folds yield the same level of performance increase. The left plot

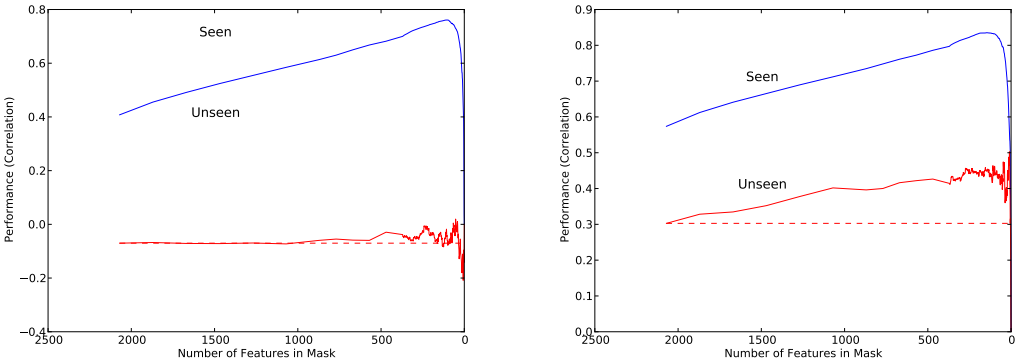

Figure 8.2: Additional examples of performance increases as feature selection progresses. The left plot shows GBR *question* performance. The right plot shows KEN *agree* performance. Both unseen data folds correspond to subject 1011.

of Figure 8.2 shows an instance in which feature selection was not effective. The performance is low before feature selection begins, which might indicate a problem with the approach in recognizing this subject performing *question* NVC signals. The right curve shows typical feature selection behaviour in a different culture (KEN in this case). A typical gradual improvement in performance can be seen, as features are removed before a sharp decline.

The optimal number of features is not known before feature selection begins. The peak of unseen performance is 10 features, while the peak for seen performance is at approximately 125 features (see Figure 8.1). A simple approach to determine the optimal number of features is to use the peak performance of unseen data. The performance for this method is shown in Table 8.1. However, this method violates the separation of training and unseen test data. Therefore, the results should not be directly compared to performance results in the previous chapter. The table shows the performance with the ideal termination of feature selection. This table implies that if terminated at an appropriate point, SBE can result in a significant performance gain.

Table 8.1: System performance with termination of features selection based on the peak of unseen performance. This violates the separation of training and test data but this shows the performance with an ideal termination point.

Area	NVC Category	Terminate By Unseen Peak
GBR	Agree	0.588
GBR	Question	0.453
GBR	Thinking	0.617
GBR	Understand	0.640
IND	Agree	0.627
IND	Question	0.534
IND	Thinking	0.638
IND	Understand	0.647
KEN	Agree	0.648
KEN	Question	0.453
KEN	Thinking	0.654
KEN	Understand	0.636
All	Average	0.586

8.2.3 Terminate SBE Based on Seen Training Data

The number of features for termination of the feature selection process should be determined based on seen training data. This restriction represents a system which is less reliant on manual tuning of parameters. The peak training data performance can be used to determine when to terminate the feature selection process. This is likely to select a non-optimal number of features, but this approach respects seen and unseen data separation. The results may be compared to the regression system in the previous chapter. The performance of this method is shown in the highlighted column of Table 8.2. Feature selection produces a large increase in performance over the method described in the previous chapter. Therefore, feature selection is beneficial for *geometric-a*

Table 8.2: Comparison of various approaches of termination of the feature selection process, along with the performance without feature selection from the previous chapter. Termination using the unseen peak, as discussed in Section 8.2.2, is the upper limit for performance based on SBE. Termination based on peak seen performance (highlighted) is discussed in Section 8.2.3.

Area	NVC Category	Terminate By Unseen Peak	Terminate By Seen Peak	Without Feature Selection
GBR	Agree	0.588	0.523	0.340
GBR	Question	0.453	0.385	0.188
GBR	Thinking	0.617	0.556	0.440
GBR	Understand	0.640	0.605	0.389
IND	Agree	0.637	0.600	0.400
IND	Question	0.534	0.458	0.236
IND	Thinking	0.638	0.588	0.363
IND	Understand	0.547	0.498	0.257
KEN	Agree	0.648	0.604	0.462
KEN	Question	0.453	0.358	0.162
KEN	Thinking	0.654	0.600	0.363
KEN	Understand	0.636	0.595	0.431
All	Average	0.586	0.531	0.336

features because it removes irrelevant features and results in a feature subset that is more suited for the specific NVC. The next section changes the feature selection folds to be person independent, rather than multi-person.

8.2.4 Terminate SBE on Person-Independent Folds

Feature selection in this chapter uses multiple folds to determine the performance as features are eliminated (Algorithm 3). Previously, each fold contained multi-person data in which every subject in the seen data was present in each feature selection

fold. This does not affect the overall system testing, which remains person independent throughout. This change may be advantageous because feature elimination would be based on a system that had been trained and tested on different subjects and may improve the generalisation. However, as feature selection is run iteratively, the system may begin to over fit the training data.

Table 8.3: Performance of the system using person independent feature selection folds (highlighted, Section 8.2.4), compared to feature selection on multi person folds (Section 8.2.3). There is little overall difference in performance.

Area	NVC Category Category	Person Independent Folds	Multi-Person Folds
GBR	Agree	0.492	0.523
GBR	Question	0.341	0.385
GBR	Thinking	0.581	0.556
GBR	Understand	0.599	0.605
IND	Agree	0.571	0.600
IND	Question	0.522	0.458
IND	Thinking	0.554	0.588
IND	Understand	0.487	0.498
KEN	Agree	0.630	0.604
KEN	Question	0.393	0.358
KEN	Thinking	0.598	0.600
KEN	Understand	0.591	0.595
All	Average	0.530	0.531

The performance of termination based on person independent folds is shown in Table 8.3. Although there are minor differences in performance for NVC categories, the overall performance is relatively unchanged. Using either of these methods, feature selection finds a subset of features that are specialised in recognition of a specific NVC signal. The next section provides a method for visualising the feature subset and an interpretation of the visualisations.

8.3 Visualising Selected Feature Subsets

Each feature component in the feature selection subset corresponds to a pair of trackers. This provides information about which facial regions are used by the regression model for NVC recognition. It is useful to know which areas of the face are involved in NVC expression: to assist understanding of human behaviour and to develop effective feature extraction methods. In order to visualise areas of the face relevant to NVC expression, each feature component of the *geometric-a* feature is assigned a weight based on the contribution that the feature component makes to the performance. As feature component i is removed at SBE iteration j, an increase $\mathbf{o}_j$ in performance from $\mathbf{p}_{j-1}$ to $\mathbf{p}_j$ where $(\mathbf{p}_j > \mathbf{p}_{j-1})$ indicates the component was detrimental and is ignored $(\mathbf{o} \in \mathbb{R}^s, \mathbf{p} \in \mathbb{R}^s)$. If a component is removed and the performance drops, this indicates the component was relevant.

$$\mathbf{o}_j = \begin{cases} |\mathbf{p}_j - \mathbf{p}_{j-1}| : \mathbf{p}_j - \mathbf{p}_{j-1} > 0 \\ 0 : \mathbf{p}_j - \mathbf{p}_{j-1} \leq 0 \end{cases} \tag{8.3}$$

The modulus of the performance drop $\mathbf{o}$ is added to the weight of the two trackers $\mathbf{P}_j^a$ and $\mathbf{P}_j^b$ that correspond to the component i $(\mathbf{P} \in \mathbb{R}^{\kappa \times s})$.

$$\mathbf{P}_{j-1}^a = \mathbf{P}_j^a + \mathbf{o}_j \tag{8.4}$$

$$\mathbf{P}_{j-1}^b = \mathbf{P}_j^b + \mathbf{o}_j \tag{8.5}$$

$$\tag{8.6}$$

After the SBE process is run to completion, the tracker weights are normalised to form normalised weight $\mathbf{n}$ which makes the tracker maximum weight equal to one $(\mathbf{n} \in \mathbb{R}^{\kappa})$.

$$\mathbf{n}^x = \frac{\mathbf{P}_{j=0}^x}{max(\mathbf{P})} \tag{8.7}$$

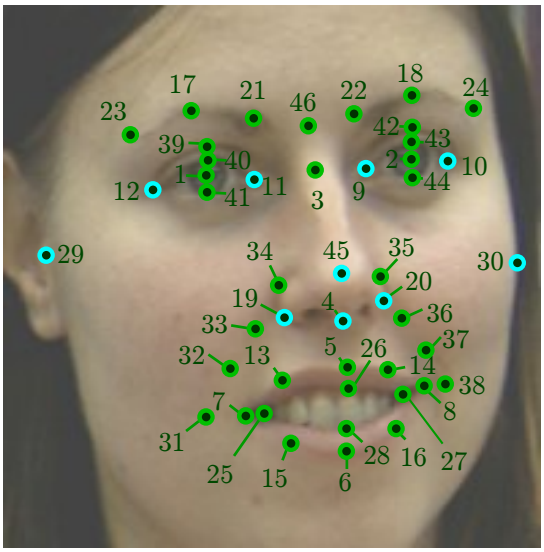

Figure 8.3: Manual division of tracker points into a flexible and rigid sets. Flexible points are shown in green. Rigid points are shown in cyan. Humans have relatively little ability to move these rigid facial points relative to the skull.

To investigate the relative importance of head pose when compared to the role expression, the trackers have been manually divided into rigid and non-rigid facial points. The manual division of trackers is shown in Figure 8.3. However, it is possible to automatically separate points into rigid and flexible sets, as described by Del Bue et al. [74]. The normalised tracker weights for each of the four NVC categories are shown in Figure 8.4. All NVC categories have significant weights assigned to trackers on flexible parts of the face, which implies expression is significant for NVC recognition. The weights assigned to rigid trackers are relatively low for *question* NVC and to some extent in *thinking*. This suggests that these NVC signals are largely conveyed by expression, with head pose having little importance. In contrast, the rigid tracker weights have higher weights in *agree*, which suggests that head pose has a role in the automatic recognition process. This confirms our expectation that agreement is often expressed by the nodding of the head. The weightings also show that the trackers that have low weightings for all of the studied NVC signals. The lowest weighted tracker overall was number 22, which corresponds to a part of the eyebrow. This may indicate a problem with this tracker or this area is redundant for recognizing these types of NVC signals.

Although each tracker weight corresponds to a specific area of the face, it is difficult to form an overall impression of which areas of the face are involved, based on these bar charts. A better approach is to visualise the relevant areas in relation to an actual face. However, the visualisation process is complicated by the head pose. Head pose changes are not localised to an specific area of the face and should be removed. The head pose is generally encoded by the distance between two rigid points on the face. Facial deformations can either be encoded by distances which are either between rigid to flexible facial points or between flexible to flexible facial points. The remaining non-rigid points correspond to the flexible regions of the face and are responsible for facial deformations. The facial areas are based on a Voronoi tessellation of the face [81], based on tracker positions on a manually selected frontal view of the face. The normalised weights of each tracked point are used to control the saturation of the local area in the image. Relevant areas are shown as normal saturation. Irrelevant areas are shown as desaturated, which makes the colour tend to pure white for low weights. This enables an intuitive way to visualise relevant areas for NVC expression around the face.

The results of the visualisation are shown in Figure 8.5. The clearest example of facial areas corresponding to our expectation is for *thinking*. The eyes are prominently selected and gaze is already thought to play a role in this NVC, as discussed in Section 4.8. The other features provide an indication into less well understood NVC. The brow region seems important for *question* NVC. When intense examples of *question* are viewed, there is generally consistent brow lowering (action unit 4) lasting for less than a second which occurs at or near the end of a question sentence (see Figure 8.6). The feature selection seems to be using this behaviour as the basis for recognition. This connection between verbal questioning and brow lowering has not been previously reported in published research, although Ekman mentions unpublished experiments which found this association [90]. Brow raising and lowering has also been documented in sign language but in this context, the direction of raising or lowering has a distinct semantic meaning, depending on the type of question that is being asked [92]. For *agree* and *understand*, the areas selected are less specific but generally indicate that the eyes and mouth are involved and the brow area is not used. While the visualisation shows areas that are involved in NVC recognition by ML, it does not necessarily imply that

humans use these areas for recognition, but shows that information is present in these areas. However, there is a strong possibility that humans also use this information during NVC perception.

This approach could be improved by additional trackers, would would improve the spatial resolution of the visualisation.

8.4 Applying Feature Selection to the Mind Reading Corpus

Classification of mental states in the Mind Reading corpus was previously discussed in Section 7.8. This section applies feature selection to the algorithmic geometric features in an attempt to improve classification performance. The motivation is the save as in the case of TwoTalk: the feature vector contains irrelevant or redundant information which can reduce the performance of recognition. The experimental arrangement used is the same as described in Section 7.8, except for only a subset of features are used, as determined by Algorithm 1. Because cross validation evaluation is performed on a leave-one-clip-out basis, it is not computationally feasible to perform a separate feature selection for each of the 174 folds. In this case, feature selection is conducted on a two fold basis on the entire corpus. As the feature selection removes features by SBE, the performance based on two fold classification accuracy rises. The change in performance during the SBE process is shown in Figure 8.7. The resultant masks are then evaluated using the standard level-one-clip-out; the classification accuracy of this is shown in Figure 8.8. As usual, the performance gradually increases as unnecessary or detrimental features are removed, before a sudden reduction in performance. Peak performance for the feature selection occurs for between 50 and 90 feature components. A set of 88 feature components is manually selected, based on it having the best classification accuracy. The number of features removed at each stage is as follows:

Table 8.4: Confusion matrix of mental state classification using geometric algorithmic features with feature selection on the Mind Reading corpus. Feature selection resulted in a set of 88 feature components which are more effective at classification.

mental state	agreeing	concentrating	disagreeing	interested	thinking	unsure	accuracy
agreeing	**20**	0	4	2	5	5	55.6%
concentrating	0	**0**	2	7	3	6	0.0%
disagreeing	7	0	**8**	2	3	4	33.3%
interested	2	2	0	**20**	2	4	66.7%
thinking	5	3	2	2	**21**	3	58.3%
unsure	1	2	2	2	5	**18**	60.0%
mean							**45.6**%

$$\eta = \begin{cases} 4 : \omega > 1000 \\ 2 : \omega > 500, \omega \leq 1000 \\ 1 : \omega \leq 500 \end{cases} \tag{8.8}$$

This removes features more slowly than in the previous section, which should help prevent useful features being eliminated. However, this results in the system being evaluated at more stages during the SBE process.

The use of feature selection results in a significant performance increase but is still not as effective as el Kaliouby and Robinson [99]. Classification for *concentrating* is worse than chance classification and is often mislabelled as *interested* or *unsure*, which are arguable closely related to *concentrating*. There are less examples of *concentrating* in the corpus, which may be a factor in poor recognition performance for this class. Other reasons for the lower performance include:

- Acted behaviour is more intense and more consistent than spontaneous behaviour. This may be exploited by the temporal model used by el Kaliouby and Robinson to achieve a better result.
- NVC and mental states are different concepts. The proposed system was developed for NVC and el Kaliouby and Robinson's system was developed to address the Mind Reading corpus.

- There is less video data available for each class and more subjects in the Mind Reading corpus, which may affect which method is suitable. Specifically, the proposed method uses subject specific normalisation which can require a significant amount of data to reach a stable and robust model of facial behaviour.
- el Kaliouby and Robinson discarded video clips that were not tracked, while we consider all videos, which is a harder problem.
- The feature extraction technique used by each method is different, which may encode relevant information that is missed by other feature extraction approaches. Although the *geometric-h* heuristic features are based on el Kaliouby and Robinson, it is not the same. They use appearance features that focus on mouth opening and teeth visible, which may be useful in distinguishing some of the classes.
- Feature selection on geometric algorithmic features was conducted based on the performance of all classes. However, it may be more effective to have a one-vs-one class basis for feature selection, to better isolate features that distinguish each class.

8.5 Conclusion

Geometric features used in the previous chapter contain a great deal of redundant and irrelevant components. This chapter describes an SBE based method to find a subset of features that are relevant for a specific NVC signal. This removes feature components that are not relevant for NVC recognition and this results in a significant performance increase. The feature subset is then visualised to show the facial areas used by the automatic system. This provides evidence of which facial areas are involved in the expression of each NVC signal. Knowing the areas of the face used for NVC can suggest feature types that better encode these local areas, avoids computation of irrelevant or redundant features, as well as improving our understanding of human behaviour.

The areas of the face that are used by the system either correspond to the expected areas, or for NVC signals that are less well understood, they give an indication as to the facial areas that are involved. The areas used for each NVC are different, which

implies that the feature selection has isolated feature components that are specific to each NVC. Thinking is known to involve gaze aversion and this is clearly seen in that feature components that encode eye movement are retained by the feature selection process. Based on reviewing corpus videos, it was manually observed that a sentence ending with a question is often accompanied by a brief brow lowering and this is also consistent with the visualisation of questioning NVC.

SBE feature selection was based on multiple folds of the seen data. Using folds that were either person-independent or person-dependent did not make a significant performance difference. The termination of the SBE process was based on the peak performance of the training data used in the optimisation. This does not select the optimal number of features but it still resulted in a significant performance increase. If a system can be manually tuned, a slightly better performance can be achieved but the number of features that is optimal depends on the specific NVC.

The visualisation of the feature selection subsets used annotation data from a single culture. It may be possible to investigate whether other cultures use different areas of the face for NVC perception, based on feature selection. Gaze patterns are culturally dependent for emotion recognition [149]. However, humans may be using different areas of the face for recognition, compared to an automatic system. This may be due to the feature extraction process not being as comprehensive as human perception. The areas used by an automatic system may provide indirect clues as to the way human perception operates. This cross cultural visualisation is not attempted in this thesis, because this would require a larger video corpus, more comprehensive facial encoding and additional annotation data to provide a reliable result.

Head pose has a role in NVC but the visualisation in this chapter is focused on local facial deformations. The features are only considered as simplistic temporal variations. The temporal encoding currently considers an entire clip, so cannot temporally localise relevant motion in NVC expression. However, with more detailed temporal encoding, which might consider variation in a sliding window, a particular time and area of the face could be identified as important for NVC automatic regression. The feature selection framework also might provide a framework to extend the existing automatic system

to other feature types. Considering many different areas of the face (or holistic facial features) over multiple time scales and temporal offsets will result in a vast number of potential features. For this reason, techniques that are suitable for spotting patterns in large data sets, such as data mining, may be relevant to facial analysis.

The feature selection method presented here is a simple but computationally intensive approach. The removal of many features during the early iterations was necessary to make the approach practical but the performance implications of this approximation are not well understood. Other feature selection methods may be investigated to reduce the computation requirements and improve performance.

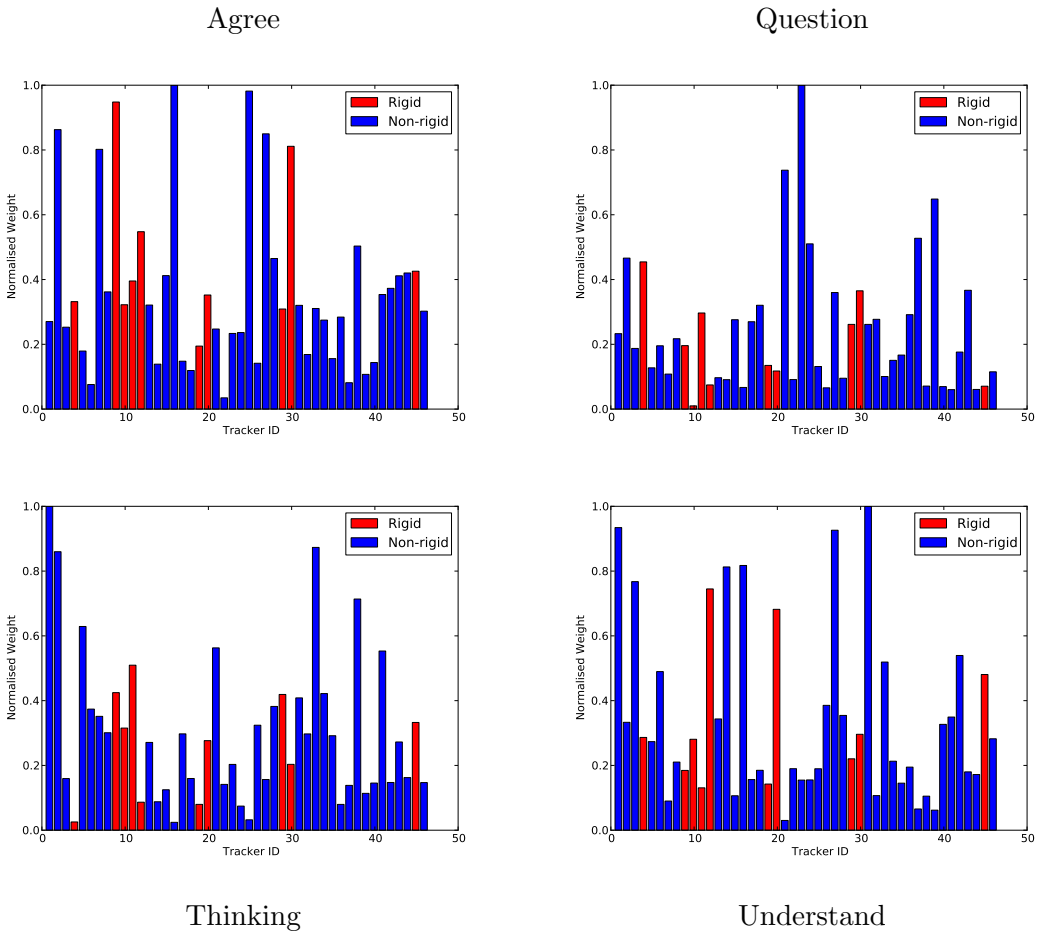

Figure 8.4: Bar charts showing the normalised weights of tracking features for the four NVC categories. Rigid and non-rigid trackers are shown as different colours, which indicate the relative importance of expression vs. head pose in recognition. The tracker ID numbers correspond to the numbering in Figure 8.3. Results are from GBR culture, with person independent folds in feature selection. Visualisation areas have been averaged across test folds.

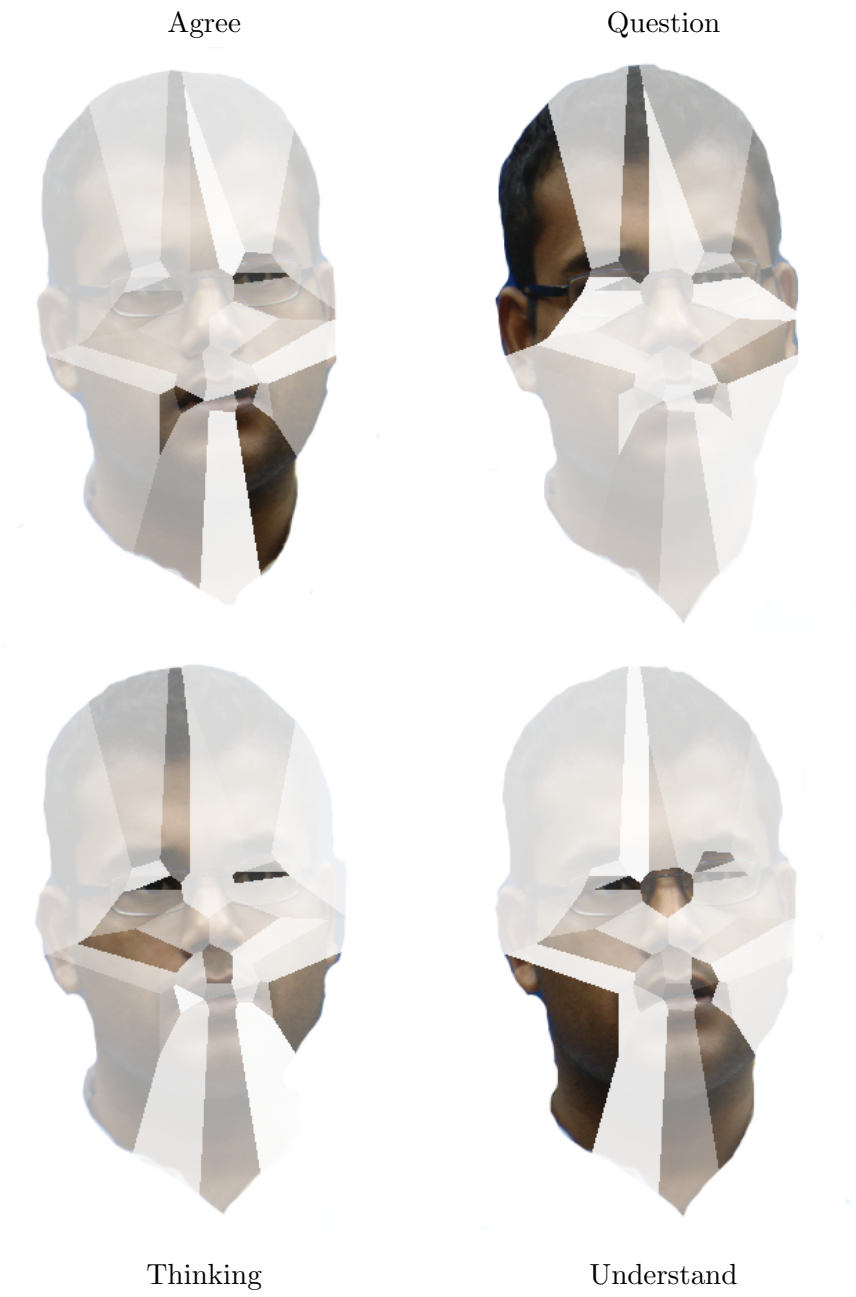

Figure 8.5: Visualising the areas of face used for feature extraction . The face is segmented based on Voronoi tessellation. More saturated areas indicate the importance of an area, less saturated areas are not relevant for a particular NVC. Results are from GBR culture, with person independent folds in feature selection. Visualisation areas have been averaged across test folds.

Figure 8.6: Brow lowering (action unit 4), lasting for less than a second, often occurs at or near the end of a question sentence.

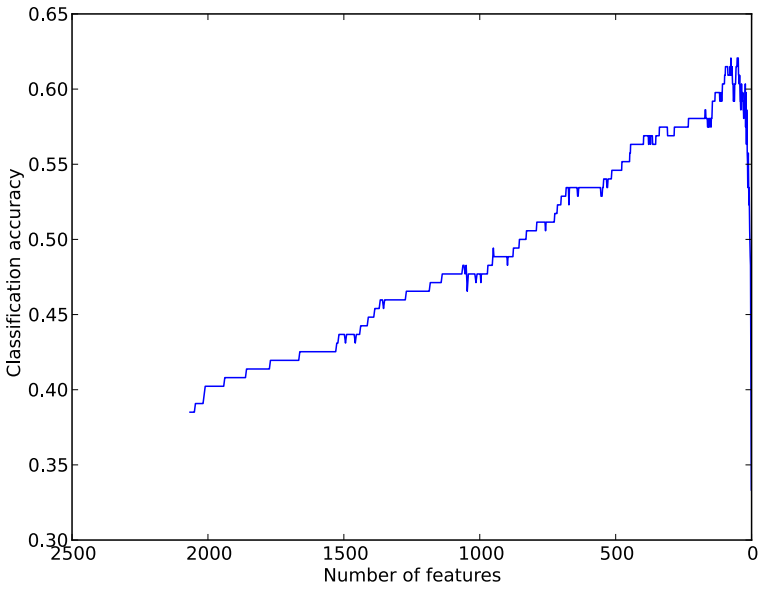

Figure 8.7: Classification accuracy of two-fold cross validation on the Mind Reading corpus as features are eliminated by SBE.

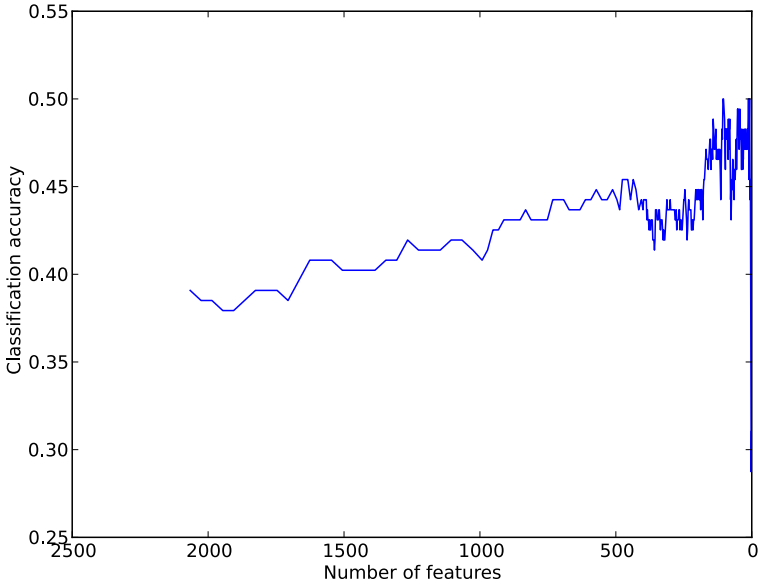

Figure 8.8: Classification accuracy of level-one-clip-out cross validation on the Mind Reading corpus as features are eliminated by SBE, based on the previously selected feature sets.

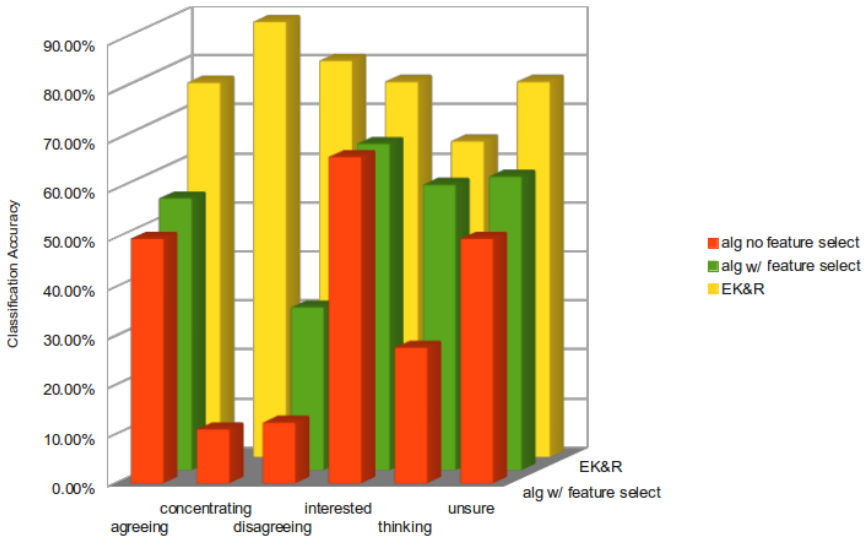

Figure 8.9: Performance of various classification methods based on the Mind Reading Corpus. Orange corresponds to algorithmic geometric features without feature selection (method described in Section 7.8). Green corresponds to the algorithmic geometric feature selection method proposed in this section, with 88 components. Yellow corresponds to el Kaliouby and Robinson [99]. Chance prediction performance is 17%.

If the fool would persist in his folly he would become wise.

William Blake

9

Conclusion and Future Directions

Intentional Non-Verbal Communication (NVC) messages are an important component of human interaction and a skill which almost all people innately possess. Automatic recognition of NVC can enable novel computer interfaces, help computers facilitate human communication and provide a tool for understanding human behaviour. Understanding human behaviour is complicated because both the expression and perception of NVC meaning is context specific. Different cultural and social situations have different rules of behaviour. For a realistic automatic system to understand social situations, the effect of context must be considered.

The main contributions of this thesis are:

1. The recording of a new corpus to study human behaviour in informal conversation. Minimal constraints were imposed on participants to allow spontaneous, natural

conversation. The corpus was made publicly available and has already been used by another research group.

2. The corpus was annotated by observers based in three different cultures. The paid annotators used an Internet based questionnaire, and their quantised dimensional responses were filtered to ensure high quality labels were produced. The labels used for annotation were based on NVC meaning. The annotation data was filtered to produce a consensus set of four component, dimensional, continuous valued of labels which encode NVC meaning based on each culture's perception of it. Various types of NVC are shown to co-occur, and others to be mutually exclusive.

3. A study of automatic NVC meaning recognition. The automatic predictions were dimensional, continuous valued labels. Various feature extraction methods and classifiers were compared. Feature selection was employed to improve performance. The temporal envelope of feature components was encoded using simple statistical measures. Culturally specific models of NVC perception were shown to be beneficial to recognition performance. Specific areas of the face were identified by the feature selection process that indicate the local areas of the face used in the automatic recognition system for specific NVC categories.

4. Coupling in human behaviour was confirmed and quantified, along with the identifying the areas of the face most relevant for NVC recognition. Analysis was performed both quantitatively and by using qualitative visualisation of the facial regions involved. The strongest coupling of behaviour was found to be associated with the mouth and possibly head pose. The use of non-corresponding features resulted in a higher level of correlation for some subjects. Classification using back channel features was shown to be above chance level, although this requires additional experiments to confirm if this finding is statistically significant.

Historically, many studies of NVC and emotion recognition have used posed data sets. However, posed human behaviour is quite different from spontaneous, natural behaviour. The corpus used in this thesis was based in a spontaneous, informal conversational setting. This context is a common social situation and it occurs in every

culture. Because NVC varies depending on social and cultural rules, recording a corpus in a specific, common situation provides a starting point for the collection of NVC data in a variety of other natural situations. Perception of NVC also depends on culture, so annotation data was collected from three different cultural groups. Differences in annotations suggests that an automatic system would need to specialise in a particular cultural view point to accurately model and recognise NVC in the manner that was comparable to a human. Annotation data was collected from paid volunteers using the Internet, a process sometimes called "crowd sourcing". This data collection method has quality problems if used directly and existing filtering methods provided by these services are only relevant for discrete, objective labels. A robust subjective consensus was created using filtering of annotators into trusted and untrusted groups.

Annotation was performed using labels that correspond to meaning in NVC. Only a few recent existing studies have attempted to recognise intentional and meaningful communication acts; the handful that do perform automatic recognition of agreement. This thesis broadens the types of NVC signals that are recognized to include asking a question, thinking and understanding NVC. The NVC of thinking, which was previously known to involve gaze aversion, could arguably be considered a mental state and not only a form of NVC. However, this gaze pattern may intentionally regulate conversational turn taking and therefore is considered a form of NVC. The annotators rated corpus samples for 4 NVC categories using quantised, dimensional labels. Most existing facial behaviour recognition studies use discrete, multi class labels. These labels ignore the fact that NVC signals and emotions occur in different intensities and these differences may be significant, depending on the intended application. Using dimensional labels retains information of the intensity of the NVC signal, which provides a much richer set of labels to use in training an automatic system.

The quantised, dimensional labels were combined to form consensus labels that were dimensional, continuous valued data. An automatic system was created and trained on the consensus annotation data provided by annotators in three distinct cultures. This enables the automatic system to better model the NVC perception of that culture. All other studies have used annotators in a single cultural group or otherwise treated them as culturally homogeneous. There are a few papers that use distinct groups of

annotators, but this has so far been used to partition annotators into naive and expert skilled annotator sets, and in which the labels are objective observations. This thesis considers annotator groups as equally valid observers of subjective behaviour meaning. Only visual data was used in training and testing the automatic system, which forces the system to use visual only NVC signals for recognition.

The automatic system evaluation compared several different approaches to determine which approach yielded the best performance. The findings were broadly consistent with existing facial analysis studies: specifically that face shape information was effective for automatic NVC recognition. The optimal features were based on distances between pairs of trackers. Combining these features with feature selection, which determines which specific pairs of trackers were important, resulted in a significant performance improvement. Some forms of NVC appear to be easier to automatically recognise than others. Even for questioning NVC, which was expected to be entirely based on audible signals, the recognition performance was relatively high. This indicates the NVC for questioning has a visual component.

Evidence of interpersonal coordination of facial behaviour was found, quantified and visualised. Movement around the mouth was correlated in people that were engaged in conversation. This is likely to be due to mutual smiling. Another coupled behaviour appears to be related to head pitch and may be related to coupled body pose or nodding of the head.

The features used by the automatic recognition system were identified by feature selection and visualised. This enables an intuitive understanding of the areas of face that were used. Previous approaches have used visualisations that are hard to interpret, such as visualising the single most significant feature. The visualisation method described was based on the overall contribution of all features that are relevant for NVC recognition. The resulting visualisations either confirm our expectations, e.g. gaze is important for thinking NVC, or suggest new insights into human behaviour, such as the brow region being significant when a question is being asked.

9.1 Current Limitations and Future Work

This thesis provides a first step in automatic understanding of NVC signals in a common social setting. However, these findings do not yet have a direct application. This is partly because they enable an entirely new type of computer interface. An interface that is NVC aware may be used in conjunction with natural language, such as interacting with computer characters using both verbal and NVC signals. Experimental systems for this already exist, such as the SEMAINE Sensitive Artificial Listener (SAL) system [279], which uses manually defined NVC rules to predict the mental state of a user and to generate appropriate back channel signals. SAL does not attempt to explicitly understand any verbal or NVC meaning in the conversation. Applying the work in this thesis would allow meaning to be explicitly understood, but it is possible to sidestep this issue and generate a computer character's response without recognizing NVC meaning. The work presented in this thesis also shows that feature extraction and selection can provide insight into human behaviour. This includes coupling of behaviour as well as identifying which facial areas are involved in NVC. Automatic systems provided evidence that relevant information is present in a specific area of the face. However, features used for recognition by automatic systems may be different from the process used by humans to recognise behaviour. This area is relatively unexplored, with computer tools only recently being applied to assist humans by automatically finding patterns in human behaviour.

There are several limitations to the work presented in this thesis. The corpus was recorded in a single social and cultural context. Having culturally distinct sets of NVC expression would allow cultural differences to be automatically analysed. This is related to Hassan's work on cross language emotion recognition based on audio [137]. Additional cultural data would enable automatic systems to be culturally specialised and therefore to improve performance. However, there are so many different cultural and social situations, that this process cannot be performed exhaustively. There are likely to be similarities within groups of social situations and this would likely reduce the scale of the problem of contextual differences. The different levels of expressivity of the conversation participants was not considered. Normalising or removing these vari-

ations, based on annotation label data, could improve person independent recognition. However, the system would then require personal adaptation to the subject's style of expression. This improvement by familiarity of personal style is commonly used by humans but is currently not considered by automatic systems. The corpus has been recorded in a laboratory environment but may modify human behaviour in comparison to behaviour in the field. A better data set might be obtained by recording in a domestic or work environment without the experimenters attempting to modify the conditions in which NVC is expressed. The latest datasets have begun to adopt this approach. However, recording NVC without the knowledge of participants in an uncontrolled environment has serious ethical concerns, particularly if the corpus is shared and used by multiple research groups for the purposes of performance benchmarking. It may be possible to use geo-centric (wearable) cameras to record human behaviour in the field but the resulting data would be challenging to process due to lighting changes, pose, the changing environment and ethical data protection issues.

Cultural differences exist in the annotation data. However, it is difficult to be sure if this is caused by perceptional differences or other factors. There are cultural differences in language usage and computer skills, as well as different samples of the population participating in annotation; these factors might also be causing changes in annotation perception. Some of these issues could be addressed by gathering demographic and personality data. Also, annotators could be grouped using unsupervised clustering and a specialised model trained in the annotator set. Further work could be done to validate the questionnaire, based on objective assessments and ensuring this is consistent across cultures.

Only four NVC signals were annotated and automatically recognized. However, there is no clear upper limit for the number of types of NVC that may be expressed. NVC can be hard to notice that it is even occurring, such as conversation turn taking where the meaning is rarely explicitly expressed. Many other NVC signals could be proposed but NVC signals are not expressed in isolation. Dimensionality reduction may enable a relatively comprehensive coverage with a finite number of NVC labels.

Many NVC signals evolve in time and cannot be accurately recognized from a photo-

graph or single frame of video. The system in this thesis encodes temporal variation by simple statistical measures for manually selected video clips. However, it is not clear how to apply this method to a real time system in which there is no clear start or end of NVC expression. A technique to select an optimal temporal window size, or another way to encode temporal information needs to be applied to create an effective real time system. The temporal encoding used in this thesis is simple. No other consistent NVC behaviours were found that required a more sophisticated encoding method, but it is quite possible these exist. Further work into identifying and temporally modelling the behaviour over several seconds (or longer) would be useful for many practical applications.

Only visual features are analysed, but NVC is closely associated with both the non-verbal aspect of voice, as well as the verbal meaning. Also, only facial features are investigated but future work might employ body pose, shoulders, hand shapes or other visual features as part of NVC. The role of clothing, proximity and other non facial NVC signals might also be useful, if automatically recognized. Future work may consider the close relationship between speech recognition and NVC recognition. Work has already been done on audio-visual speech recognition but little has been done on multi-modal NVC recognition. Combining NVC with verbal meaning may result in an increase in verbal recognition accuracy due to the close association between words, emotions and NVC.

Tracking is performed using a method that requires annotation of multiple frames and occasionally needs manual re-initialisation after occlusions. This is not suitable for a production system that operates on previously unseen people. A system that uses facial detection to reinitialise would be more robust to occlusions. Although the tracker is relatively robust to limited pose change, large pose change still causes significant tracker drift due to the radical change in appearance. Large pose changes would need to be tolerated by a system that needs to operate in situations in which users move around freely.

This thesis has used various methods to recognise NVC in natural data. Many other techniques exist, particularly for emotion recognition. However, most existing tech-

niques have only been applied to posed data. It would be informative to see which techniques generalise from posed data to natural situations. This is a necessary step if a technique is to be applied to applications in an unconstrained environment. In particular, there are many feature extraction methods. It is still unclear if shape or appearance is more useful for automatic facial analysis in general but previous studies have concluded that fusion of multiple strategies is likely to provide optimal performance. The current geometric features are distance based and are sensitive to scale changes. These features may be generalised to include other types of geometric arrangements, as well as made robust to scale and pose change.

The annotation was performed by observers using Internet based tools. The introduction of this technique has already had an impact in the availability of training data in many fields, such as object recognition, and it is beginning to be applied to human behaviour. However quality issues are still a problem as some workers do not cooperate with the task. This issue is difficult to address for subjective questions or if the responses are based on dimensional, continuous valued scales. Given recent advances in crowd sourcing services, it should be possible to target questionnaires to particular regions and perhaps to gather demographic and personality based information. This will provide a much richer understanding of perception differences based on context.

Research based on a practical application of NVC recognition may provide a set of constraints for the problem, as well as providing clues as to the most suitable environment for recording and which labels are appropriate for annotation. The XBox Kinect has provided an application for body pose estimation, and with it a set of assumptions for the hardware platform, distance from the camera and domestic operating environment. In a similar way, a concrete application for NVC may provide guidance as to suitable constraints and assumptions that can be used, as well as the factors that require system robustness.

Automatic NVC recognition remains a relatively new and open area of enquiry. It is also fragmented between various academic disciplines, each with their own practices, jargon and methodologies. The field is likely to benefit from exchange of ideas from these different groups, including between computer vision, psychology, anthropology,

linguistics and engineering. However, the goal of socially aware computing in every-day life is still some way away.

A

Additional Classification Results

Table A.1: Adaboost, Multi-person testing, classification of sliding window examples. Mean and standard deviation performance is shown.

Test	Agree	Question	Think	Understand
affine	0.62±0.09	0.52±0.03	0.53±0.06	0.64±0.03
deform-cubica	0.51±0.11	0.60±0.04	0.57±0.03	0.51±0.06
deform-fastica	0.51±0.11	0.60±0.04	0.57±0.03	0.51±0.06
deform-pca	0.59±0.07	0.65±0.05	0.70±0.06	0.65±0.05
geometric-h	0.64±0.09	0.69±0.04	0.60±0.07	0.65±0.05
geometric-a	0.65±0.03	0.62±0.07	0.67±0.03	0.71±0.01
lbp	0.57±0.08	0.67±0.03	0.57±0.04	0.59±0.04
lm	0.56±0.11	0.66±0.03	0.59±0.04	0.46±0.05
affine-t	0.62±0.09	0.52±0.03	0.53±0.06	0.64±0.03
deform-cubica-t	0.52±0.12	0.60±0.04	0.58±0.03	0.51±0.06
deform-fastica-t	0.46±0.13	0.61±0.05	0.58±0.03	0.49±0.06
deform-pca-t	0.59±0.07	0.65±0.05	0.70±0.06	0.65±0.05
geometric-h-t	0.64±0.09	0.69±0.04	0.60±0.07	0.65±0.05
lm-t	0.56±0.11	0.66±0.03	0.60±0.03	0.46±0.05

Table A.2: Adaboost, Multi-person testing, classification of video clips

Test	Agree	Question	Think	Understand
affine	0.59±0.08	0.54±0.03	0.51±0.03	0.60±0.02
deform-cubica	0.50±0.11	0.61±0.05	0.59±0.02	0.49±0.04
deform-fastica	0.50±0.11	0.61±0.05	0.59±0.02	0.49±0.04
deform-pca	0.61±0.05	0.68±0.03	0.70±0.07	0.72±0.05
geometric-h	0.64±0.09	0.71±0.04	0.66±0.08	0.69±0.04
geometric-a	0.70±0.03	0.68±0.11	0.75±0.01	0.75±0.01
lbp	0.58±0.06	0.68±0.04	0.60±0.06	0.62±0.05
lm	0.55±0.06	0.65±0.02	0.59±0.03	0.46±0.05
affine-t	0.59±0.08	0.54±0.03	0.53±0.05	0.59±0.03
deform-cubica-t	0.51±0.12	0.61±0.05	0.59±0.02	0.49±0.04
deform-fastica-t	0.45±0.14	0.64±0.05	0.60±0.00	0.48±0.04
deform-pca-t	0.61±0.05	0.68±0.03	0.70±0.07	0.72±0.05
geometric-h-t	0.64±0.07	0.71±0.04	0.66±0.08	0.69±0.03
lm-t	0.55±0.06	0.65±0.02	0.60±0.02	0.46±0.05

Table A.3: SVM, Multi-person testing, classification of sliding window examples

Test	Agree	Question	Think	Understand
affine	0.47±0.03	0.52±0.02	0.48±0.04	0.51±0.01
deform-cubica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca	0.50±0.02	0.54±0.02	0.53±0.03	0.54±0.01
geometric-h	0.58±0.04	0.57±0.02	0.61±0.01	0.63±0.02
geometric-a	0.66±0.06	0.67±0.05	0.74±0.01	0.73±0.03
lbp	0.57±0.04	0.58±0.04	0.55±0.02	0.58±0.07
lm	0.50±0.03	0.58±0.03	0.56±0.04	0.50±0.01
affine-t	0.47±0.04	0.52±0.03	0.48±0.04	0.52±0.01
deform-cubica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca-t	0.50±0.04	0.61±0.03	0.55±0.06	0.54±0.02
geometric-h-t	0.60±0.04	0.57±0.03	0.63±0.02	0.64±0.02
lm-t	0.51±0.04	0.60±0.02	0.59±0.04	0.50±0.03

Table A.4: SVM, Multi person testing, classification of video clips

Test	Agree	Question	Think	Understand
affine	0.49±0.05	0.56±0.03	0.49±0.07	0.54±0.05
deform-cubica	0.51±0.01	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica	0.51±0.01	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca	0.50±0.02	0.61±0.02	0.56±0.07	0.49±0.03
geometric-h	0.71±0.05	0.69±0.04	0.73±0.03	0.77±0.03
geometric-a	0.70±0.06	0.70±0.06	0.83±0.01	0.75±0.03
lbp	0.57±0.05	0.60±0.03	0.54±0.03	0.59±0.09
lm	0.50±0.08	0.58±0.07	0.57±0.04	0.45±0.03
affine-t	0.49±0.06	0.55±0.01	0.47±0.07	0.55±0.05
deform-cubica-t	0.51±0.01	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica-t	0.51±0.01	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca-t	0.55±0.02	0.65±0.03	0.57±0.06	0.55±0.04
geometric-h-t	0.71±0.04	0.67±0.04	0.73±0.05	0.76±0.04
lm-t	0.53±0.09	0.67±0.07	0.60±0.06	0.47±0.04

Table A.5: Adaboost, Person independent testing, classification of sliding window examples

Test	Agree	Question	Think	Understand
affine	0.61±0.02	0.42±0.09	0.50±0.03	0.55±0.04
deform-cubica	0.55±0.07	0.52±0.08	0.55±0.06	0.45±0.04
deform-fastica	0.55±0.07	0.52±0.08	0.55±0.06	0.45±0.04
deform-pca	0.61±0.06	0.47±0.04	0.64±0.03	0.57±0.02
geometric-h	0.66±0.06	0.61±0.06	0.52±0.06	0.56±0.06
geometric-a	0.66±0.02	0.52±0.04	0.65±0.02	0.68±0.05
lbp	0.57±0.04	0.53±0.13	0.49±0.07	0.44±0.05
lm	0.56±0.03	0.54±0.10	0.50±0.14	0.41±0.04
affine-t	0.62±0.02	0.47±0.04	0.49±0.03	0.55±0.03
deform-cubica-t	0.59±0.01	0.56±0.06	0.51±0.02	0.47±0.01
deform-fastica-t	0.55±0.07	0.52±0.08	0.55±0.06	0.45±0.04
deform-pca-t	0.61±0.06	0.47±0.04	0.64±0.03	0.57±0.02
geometric-h-t	0.66±0.06	0.61±0.06	0.53±0.08	0.56±0.06
lm-t	0.57±0.03	0.56±0.09	0.50±0.14	0.41±0.04

Table A.6: Adaboost, Person independent testing, classification of video clips

Test	Agree	Question	Think	Understand
affine	0.62±0.05	0.39±0.07	0.50±0.03	0.53±0.09
deform-cubica	0.52±0.05	0.56±0.11	0.55±0.07	0.42±0.04
deform-fastica	0.52±0.05	0.56±0.11	0.55±0.07	0.42±0.04
deform-pca	0.68±0.05	0.50±0.09	0.69±0.08	0.60±0.04
geometric-h	0.65±0.07	0.64±0.07	0.55±0.09	0.57±0.08
geometric-a	0.70±0.03	0.58±0.08	0.74±0.03	0.71±0.06
lbp	0.57±0.07	0.49±0.16	0.48±0.09	0.39±0.08
lm	0.56±0.02	0.51±0.10	0.48±0.15	0.46±0.04
affine-t	0.65±0.04	0.39±0.07	0.49±0.02	0.58±0.05
deform-cubica-t	0.56±0.01	0.62±0.07	0.52±0.04	0.45±0.03
deform-fastica-t	0.53±0.05	0.56±0.11	0.55±0.07	0.43±0.05
deform-pca-t	0.68±0.05	0.50±0.09	0.69±0.08	0.60±0.04
geometric-h-t	0.65±0.07	0.64±0.07	0.55±0.10	0.56±0.09
lm-t	0.56±0.02	0.54±0.07	0.46±0.13	0.44±0.04

Table A.7: SVM, Person independent testing, classification of sliding window examples.

Test	Agree	Question	Think	Understand
affine	0.49±0.03	0.51±0.04	0.45±0.04	0.50±0.03
deform-cubica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
geometric-h	0.59±0.02	0.51±0.02	0.52±0.02	0.56±0.02
geometric-a	0.66±0.03	0.54±0.01	0.70±0.03	0.70±0.02
lbp	0.55±0.05	0.51±0.11	0.49±0.08	0.47±0.08
lm	0.50±0.00	0.51±0.01	0.49±0.01	0.51±0.00
affine-t	0.49±0.03	0.52±0.04	0.46±0.04	0.50±0.03
deform-cubica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca-t	0.50±0.00	0.51±0.01	0.50±0.00	0.50±0.00
geometric-h-t	0.60±0.02	0.51±0.02	0.55±0.02	0.58±0.02
lm-t	0.51±0.00	0.51±0.01	0.48±0.01	0.50±0.01

Table A.8: SVM, Person independent testing, classification of video clips.

Test	Agree	Question	Think	Understand
affine	0.53±0.05	0.56±0.07	0.48±0.02	0.56±0.04
deform-cubica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
geometric-h	0.73±0.04	0.54±0.08	0.58±0.00	0.68±0.08
geometric-a	0.67±0.02	0.64±0.05	0.77±0.04	0.71±0.05
lbp	0.60±0.02	0.51±0.12	0.47±0.07	0.49±0.06
lm	0.50±0.02	0.50±0.03	0.47±0.01	0.49±0.01
affine-t	0.55±0.04	0.55±0.06	0.49±0.03	0.54±0.07
deform-cubica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca-t	0.50±0.01	0.50±0.01	0.50±0.01	0.50±0.00
geometric-h-t	0.74±0.03	0.57±0.05	0.60±0.02	0.69±0.06
lm-t	0.50±0.03	0.50±0.06	0.44±0.01	0.47±0.03

A.1 Performance of NVC Using Various Backchannel Features and Classifiers

Row averages from Tables A.9 to A.12 were used to generate Table 5.4.

Table A.9: Boost, multi-person testing, classification of video clips, receiver features

Test	Agree	Question	Think	Understand
affine	0.46±0.10	0.63±0.06	0.53±0.07	0.49±0.04
deform-cubica	0.44±0.06	0.67±0.03	0.51±0.02	0.45±0.07
deform-fastica	0.44±0.06	0.67±0.03	0.51±0.02	0.45±0.07
deform-pca	0.51±0.02	0.70±0.04	0.68±0.02	0.59±0.02
geometric-h	0.58±0.03	0.69±0.01	0.73±0.05	0.62±0.04
geometric-a	0.67±0.03	0.62±0.05	0.72±0.06	0.70±0.04
lbp	0.51±0.05	0.69±0.04	0.58±0.05	0.54±0.01
lm	0.48±0.06	0.68±0.04	0.55±0.05	0.51±0.09
affine-t	0.46±0.10	0.63±0.06	0.52±0.06	0.47±0.06
deform-cubica-t	0.45±0.05	0.67±0.03	0.51±0.02	0.47±0.07
deform-fastica-t	0.45±0.05	0.67±0.03	0.51±0.02	0.47±0.07
deform-pca-t	0.51±0.02	0.70±0.04	0.68±0.02	0.59±0.02
geometric-h-t	0.58±0.03	0.69±0.01	0.73±0.05	0.61±0.04
lm-t	0.48±0.06	0.68±0.04	0.55±0.05	0.48±0.08

Table A.10: Boost, person-independent testing, classification of video clips, receiver features

Test	Agree	Question	Think	Understand
affine	0.43±0.05	0.57±0.05	0.53±0.02	0.46±0.04
deform-cubica	0.41±0.03	0.60±0.06	0.45±0.08	0.51±0.04
deform-fastica	0.41±0.03	0.60±0.06	0.45±0.08	0.51±0.04
deform-pca	0.48±0.02	0.60±0.07	0.47±0.04	0.53±0.07
geometric-h	0.61±0.06	0.57±0.11	0.59±0.06	0.58±0.04
geometric-a	0.63±0.04	0.43±0.04	0.65±0.13	0.64±0.06
lbp	0.46±0.06	0.54±0.13	0.45±0.10	0.45±0.06
lm	0.45±0.07	0.64±0.03	0.42±0.04	0.49±0.07
affine-t	0.44±0.04	0.52±0.10	0.49±0.04	0.45±0.05
deform-cubica-t	0.41±0.02	0.60±0.06	0.45±0.08	0.51±0.03
deform-fastica-t	0.40±0.02	0.62±0.03	0.47±0.10	0.52±0.03
deform-pca-t	0.48±0.02	0.60±0.07	0.47±0.04	0.53±0.07
geometric-h-t	0.61±0.06	0.57±0.11	0.59±0.06	0.54±0.03
lm-t	0.45±0.07	0.64±0.03	0.42±0.05	0.49±0.06

Table A.11: SVM, multi-person testing, classification of video clips, receiver features

Test	Agree	Question	Think	Understand
affine	0.46±0.05	0.59±0.02	0.47±0.05	0.46±0.03
deform-cubica	0.51±0.01	0.50±0.00	0.51±0.01	0.50±0.00
deform-fastica	0.51±0.01	0.50±0.00	0.51±0.01	0.50±0.00
deform-pca	0.53±0.02	0.59±0.04	0.52±0.06	0.51±0.02
geometric-h	0.64±0.05	0.69±0.03	0.65±0.04	0.63±0.06
geometric-a	0.67±0.04	0.58±0.03	0.72±0.04	0.71±0.02
lbp	0.61±0.01	0.59±0.05	0.51±0.07	0.46±0.05
lm	0.54±0.02	0.63±0.02	0.53±0.03	0.50±0.05
affine-t	0.45±0.05	0.57±0.02	0.47±0.06	0.45±0.02
deform-cubica-t	0.51±0.01	0.50±0.00	0.51±0.01	0.50±0.00
deform-fastica-t	0.51±0.01	0.50±0.00	0.51±0.01	0.50±0.00
deform-pca-t	0.52±0.03	0.62±0.04	0.55±0.05	0.53±0.03
geometric-h-t	0.65±0.04	0.70±0.04	0.66±0.04	0.61±0.04
lm-t	0.56±0.05	0.66±0.03	0.54±0.04	0.50±0.07

Table A.12: SVM, person-independent testing, classification of video clips, receiver features

Test	Agree	Question	Think	Understand
affine	0.43±0.03	0.57±0.07	0.46±0.07	0.48±0.04
deform-cubica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
geometric-h	0.59±0.04	0.55±0.04	0.49±0.03	0.54±0.03
geometric-a	0.64±0.07	0.43±0.06	0.60±0.12	0.67±0.06
lbp	0.54±0.07	0.51±0.11	0.43±0.05	0.47±0.06
lm	0.47±0.04	0.49±0.02	0.48±0.01	0.48±0.01
affine-t	0.41±0.05	0.57±0.08	0.46±0.08	0.48±0.05
deform-cubica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-fastica-t	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00
deform-pca-t	0.49±0.01	0.50±0.01	0.50±0.00	0.50±0.01
geometric-h-t	0.62±0.05	0.55±0.09	0.54±0.02	0.56±0.03
lm-t	0.47±0.04	0.47±0.05	0.46±0.02	0.45±0.04

B

Linear Predictor Feature Tracking

Natural conversations contain wide range of head poses, which may be subject to sudden and rapid pose change. This work uses a pre-existing tracking method proposed by Ong et al. [231], called LP feature tracking. This section introduces the theory of LP trackers because it has not previously been applied to emotion or NVC recognition.

An LP tracker is associated with m "support pixels", where $m \geq 1$. Each support pixel has a fixed 2D offset $\mathbf{c}$ from the LP tracker position. Support pixels surrounding the LP position is depicted in Figure B.1. Also, each support pixel is assigned an initial pixel intensity at training time. Point trackers are required to predict a point of interest's motion in a series of frames. This can be expressed as the tracker providing a prediction to move from the position on the previous frame $\mathbf{T}^{n-1}$ to the position of interest's location on the current frame $\mathbf{T}^{n}$. An LP makes the motion prediction based on the difference $\delta\mathbf{p}$ between the initial support pixel intensities $\mathbf{V}^{I}$ and the support

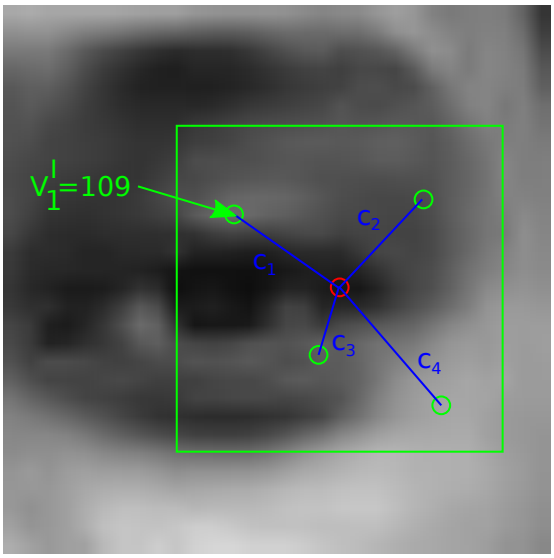

Figure B.1: The basic structure of an LP is shown, with the position of interest marked in red. There are four support pixels with offsets, $\mathbf{c}_1$ to $\mathbf{c}_4$. The image intensity for support pixel 1 is shown ($\mathbf{V}_1^I$). The square shows the area from which the support pixels are sampled.

pixel intensities on the current frame $\mathbf{V}$. Typically, images are converted to grey-scale, since the hue of the face is approximately uniform and therefore not useful for tracking. The predicted motion from the previous frame to the current frame $\mathbf{t}$ is a simple linear mapping $\mathbf{H}$:

$$\mathbf{t} = \mathbf{H}\delta\mathbf{p} \tag{B.1}$$

$$\delta\mathbf{p} = \mathbf{V} - \mathbf{V}^I \tag{B.2}$$

Training an LP is based on one or more frames in which the position of interest has been manually specified. Multiple frames are used because it enables the LP tracker to generalise to multiple head poses. m support pixel offsets are uniformly sampled from a square centred on the tracker position. The initial support pixel intensities $\mathbf{V}^I$ are set to the average pixel intensity of the support pixel locations in the training frames. Training examples are generated synthetically, based on offset the tracker by a known offset $\mathbf{t}$ and storing the pixel intensity differences at the offset position. Many synthetic

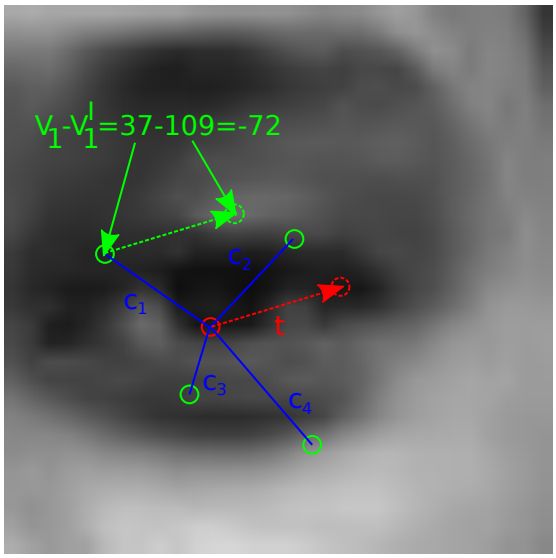

Figure B.2: The LP is offset by $\mathbf{t}$, which results in changes in intensity for all the support pixels. The intensity change for support pixel 1 is shown. The object of linear predictors is to find a linear mapping $\mathbf{H}$ between position offset and the change in support pixel intensities.

offsets can be generated for each training frame, resulting in h. The training intensity differences stored in a matrix $\mathbf{T}$ of size h by m. The training offset matrix $\delta\mathbf{P}$ is of size h by 2. The mapping $\mathbf{H}$ can be found by taking the pseudo-inverse of the training intensities.

$$\mathbf{H} = \delta\mathbf{P}\mathbf{T}^{+} \tag{B.3}$$

Applying LPs to an unseen sequence involves finding the pixel intensities $\mathbf{V}$ at the current tracker position and applying equations B.1 and B.2 to find the predicted motion.

One common failure mode of the tracking is for a single tracker to drift, while the others track correctly. These errors can often be corrected by using shape constraints, which restrict face deformations to be similar to previously observed configurations. This was a common technique that provides the basis for Active Shape Model (ASM)s [61]. The technique is applied by learning a PCA shape model from the training frames and then

modifying the feature tracking procedure, so each frame is processed as follows:

1. track new frame based on position from previous frame,
2. project tracking positions into PCA space,
3. retain the first r eigenvalues and discard the rest,
4. reconstruct the tracking positions, then
5. perform tracking as further refinement, starting at the reconstructed tracker positions.

The number of eigenvalues to retain r is normally selected to preserve 90% (or some other manually specified proportion) of the variation of the training data.

Section 4.2.1 describes how this method is applied to NVC classification.

C

Photograph Permissions

Photos used in Figure 1.1:

"Waving goodbye after the ceremony" ©Jonathan Elderfield,
https://secure.flickr.com/photos/heatheranneulrich/3420075962/
Used by permission, under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 2.0 Generic License
https://creativecommons.org/licenses/by-nc-nd/2.0/

"Bank of America security trying to prevent me from taking a photo during the Iraq war protest" ©Steve Rhodes
https://secure.flickr.com/photos/ari/2347593532/
Used by permission, under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic License
https://creativecommons.org/licenses/by-nc-sa/2.0/

Photos used in Figure 1.2:

"Eye contact" ©Jessie Reeder

https://secure.flickr.com/photos/elizacole/2503079623/

Used by permission, under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 2.0 Generic License

https://creativecommons.org/licenses/by-nc-nd/2.0/

"Wink" ©Dave77459

https://secure.flickr.com/photos/dave77459/3083102723/

Used by permission, under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic License

https://creativecommons.org/licenses/by-nc-sa/2.0/

D

Questionnaire

Rate Communication Signals in Human Conversation

Instructions

View the video sequences with audio turn on and answer the following questions.

Videos are displayed in this page (for your convenience) if you use a browser that supports the latest video standards (Firefox 3.5, etc) or Flash.

Does this person disagree or agree with what is being said?(required)

A score of 5 is neutral or not applicable.

Is this person thinking hard?(required)

Is this person asking a question?(required)

Is this person indicating they understand what is being said to them?(required)

E

Behaviour Mimicry and Pearson's Correlation

Section 5.2 describes a method for automatically identifying some types of behaviour including some types of synchrony and mimicry. Detecting behaviour by comparing simulations frames recorded on two cameras of two different subjects may seem counter intuitive because mimicry occurs in response to another behaviour. Although it is true that samples are paired in terms of observation time, the correlation coefficient is based on the overall video. This allows coupled variations to be identified, particularly if they are slowly varying.

A simple synthetic illustrative example of mimicry will now be discussed to demonstrate that Pearson's correlation is sensitive to mimicry in this case. Imagine two people in conversation with their behaviours being monitored. Assume we have a single behaviour measure for each person; what this represents is totally arbitrary but it could be some measure of body pose or facial expression. In a simple case, one person repeatedly

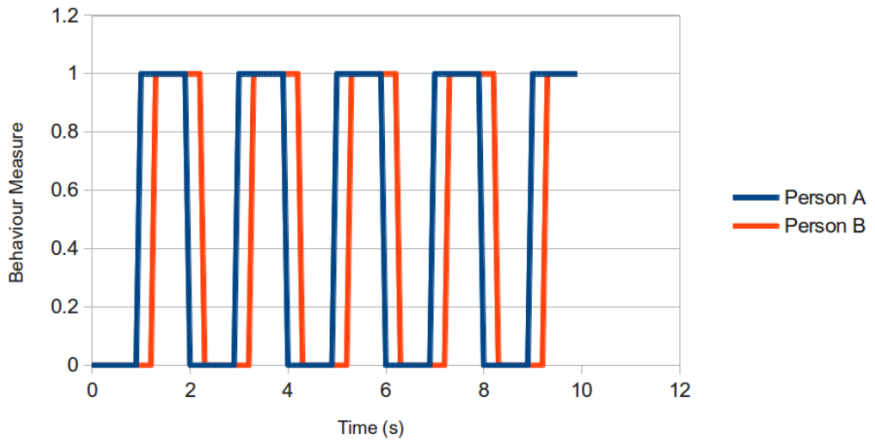

Figure E.1: A synthetic example of mimicry: person A changes their behaviour at 1Hz (the period is 0.5Hz). The time delay of person B's mimicked behaviour is 0.3 sec.

changes behaviour and the other person mimics this after a time delay. For simplicity, person A changes their behaviour at 1Hz (the period is 0.5Hz). The time delay of person B's mimicked behaviour to be 0.3 sec. Of course, most behaviours are not this regular or rapid but this is just a simple illustration of the principle. The variation of the behaviour is shown in Figure E.1.

When using the method proposed in Section 5.2, the Pearson correlation of these two components is 0.46. This demonstrates that for this type of mimicry, the method can find measurable linear coupling between these two variables. As long as there is some overlap in behaviours, the approach is sensitive and therefore suitable. This method is not suitable for certain types of rapid behaviours. For example, if blinking was coupled, this approach would not be suitable because the behaviour of Person A would likely have ended before Person B would have started to mimic it.

Bibliography

[1] AMI project website. http://corpus.amiproject.org/.

[2] Merriam Webster Dictionary. http://www.merriam-webster.com.

[3] The triumph of English. *The Economist*, 20th Dec 2001.

[4] Human gestures perplex Asimo, Honda museum robot guide. *BBC*, 5th July 2013. http://www.bbc.co.uk/news/technology-23196867.

[5] Aaron and Ben-Ze'ev. The affective realm. *New Ideas in Psychology*, 15(3):247 – 259, 1997.

[6] S. Abrilian, L. Devillers, and J. Martin. Annotation of emotions in real-life video interviews: Variability between coders. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006.

[7] S. Afzal and P. Robinson. Natural affect data - collection & annotation in a learning context. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[8] S. Afzal, T.M. Sezgin, G. Gao, and P. Robinson. Perception of the emotional expressions in different representations using facial feature points. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[9] J. Ahlberg. Candide-3 – an updated parameterized face. Technical report, Dept. of Electrical Engineering, Linkping University, Sweden, 2001. Report No. LiTH-ISY-R-2326.

[10] H. Ç. Akakın and B. Sankur. Robust classification of face and head gestures in video. *Image and Vision Computing*, 29:470–483, June 2011.

[11] Peter A. Andersen. When one cannot not communicate: A challenge to motley's traditional communication postulates. *Communication Studies*, 42(4):309–325, 1991.

[12] O. Aran, İ. Ari, M. A. Güvensan, H. Haberdar, Türkmen Kurt, Z., A. H. İ., Uyar, and L. Akarun. A database of non-manual signs in Turkish sign language. In *Proceedings of the IEEE 15th Signal Processing and Communications Applications Conference*, Eskişehir, 2007.

[13] Oya Aran, Hayley Hung, and Daniel Gatica-Perez. A multimodal corpus for studying dominance in small group conversations. In *LREC workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, Malta, May 2010.

[14] D. Archer and R. M. Akert. Words and everything else: Verbal and nonverbal cues in social interpretation. *Journal of Personality and Social Psychology*, 35:443–449, 1977.

[15] M. Argyle and M. Cook. *Gaze and Mutual Gaze*. Cambridge University Press, New York, 1976.

[16] Michael Argyle. *Non-Verbal Communication*, chapter Non-verbal Communication in Human Social Interaction. Cambridge University Press, 1975.

[17] Michael Argyle. *The Psychology of Interpersonal Behaviour*. Penguin, 5th edition edition, 1994.

[18] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn, Tsuhan Chen, Zara Ambadar, Ken Prkachin, Patty Solomon, and Barry J. Theobald. The painful face: pain expression recognition using active appearance models. In *Proceedings of the 9th International Conference on Multimodal interfaces*, pages 9–14, New York, NY, USA, 2007. ACM.

[19] A. Asthana, T.K. Marks, M.J. Jones, K.H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 937 –944, nov. 2011.

[20] Niels G. Waller Auke Tellegen. *SAGE handbook of personality theory and assessment: Volume 2 Personality measurement and testing.*, chapter Exploring Personality Through Test Construction: Development of the Multidimensional Personality Questionnaire, pages 261–293. SAGE Publications Ltd, 2008.

[21] S.E. Avons, R.G. Leiser, and D.J. Carr. Paralanguage and human-computer interaction. part 1: identification of recorded vocal segregates. *Behaviour & Information Technology*, 8(1):13–21, 1989.

[22] B Azar. What's in a face? *Monitor on Psychology*, 31(1), Jan 2000.

[23] Gene Ball and Jack Breese. Relating personality and behavior: Posture and gestures. In Ana Paiva, editor, *Affective Interactions*, volume 1814 of *Lecture Notes in Computer Science*, pages 196–203. Springer Berlin Heidelberg, 2000.

[24] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. 3D constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR 2012)*, Providence, RI, June 2012.

[25] Elizabeth Barnett and Michele Casper. A definition of social environment. *American Journal of Public Health*, 91(3), March 2001.

[26] M. Bartlett and J. Whitehill. *Handbook of Face Perception*, chapter Automated facial expression measurement: Recent applications to basic research in human behavior, learning, and education. Oxford University Press., 2010.

[27] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proceedings of the 7th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 223–230, Washington, DC, USA, 2006. IEEE Computer Society.

[28] J. B. Bavelas and N. Chovil. Faces in dialogue. In J. A. Russell and J. M. Fernandez-Dols, editors, *The Psychology of Facial Expression*, pages 334–346. Cambridge University Press, Cambridge, UK, 1997.

[29] D. E. Beaton, C. Bombardier, F. Guillemin, and M. B. Ferraz. Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine*, 25(24):3186–3191, 2000.

[30] G. Beattie. Unnatural behaviour in the laboratory. *New Scientist*, 96:181, 1982.

[31] C. Becker-Asano. Invited commentary: On guiding the design of an ill-defined phenomenon. *International Journal of Synthetic Emotions*, 2(2):66–67, 2011.

[32] Daniel Bernhardt and Peter Robinson. Detecting affect from non-stylised body motions. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, pages 59–70, Lisbon, Portugal, September 2007.

[33] F. Bernieri and R. Rosenthal. *Fundamentals of nonverbal behavior. Studies in emotion and social interaction*, chapter Interpersonal coordination: Behavior matching and interactional synchrony, pages 401–432. Cambridge University Press, New York, 1991.

[34] Bernd Bickel, Manuel Lang, Mario Botsch, Miguel A. Otaduy, and Markus Gross. Pose-space animation and transfer of facial details. In *Proc. of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pages 57–66, jul 2008.

[35] T. Blaschke and L. Wiskott. CuBICA: Independent component analysis by simultaneous third- and fourth-order cumulant diagonalization. *IEEE Transactions on Signal Processing*, 52(5):1250–1256, May 2004.

[36] Patrick Bourgeois and Ursula Hess. The impact of social context on mimicry. *Biological Psychology*, 77(3):343 – 352, 2008.

[37] K. Bousmalis, M. Mehu, and M. Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, volume 2, Amsterdam, Netherlands, September 2009.

[38] K. Bousmalis, L.-P. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2011.

[39] K. Bousmalis, S. Zafeiriou, L.-P. Morency, and M. Pantic. Infinite hidden conditional random fields for human behavior analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2012.

[40] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.

[41] Leo Breiman. Random forests. *Machine Learning*, 45(1):532, 2001.

[42] Bonnie Brinton, Matthew P. Spackman, Martin Fujiki, and Jenny Ricks. What should chris say? the ability of children with specific language impairment to recognize the need to dissemble emotions in social situations. *J Speech Lang Hear Res*, 50(3):798–811, 2007.

[43] Martin Brüne, Mona Abdel-Hamid, Claudia Sonntag, Caroline Lehmkämper, and Robyn Langdon. Linking social cognition with social interaction: Non-verbal expressivity, social competence and "mentalising" in patients with schizophrenia spectrum disorders. *Behavioral and Brain Functions*, 5(6), 2009.

[44] Ioan Buciu, Costas Kotropoulos, and Ioannis Pitas. Comparison of ica approaches for facial expression recognition. *Signal, Image and Video Processing*, 3(4):345–361, 2009.

[45] R Buck. Measuring individual differences in the nonverbal communication of affect: The slide-viewing paradigm. *Human Communication Research*, 6:47–57, 1979.

[46] Buller D. B. & Woodall W. G. Burgoon, J. K. *Nonverbal communication: The unspoken dialogue*. Harper & Row, New York, 1996.

[47] C. Busso and S.S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: A single subject study. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2331 –2347, nov. 2007.

[48] N. Campbell and A. Tabeta. A software toolkit for viewing annotated multimodal data interactively over the web. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2010.

[49] James Carifio and Rocco J. Perla. Ten common misunderstandings, misconceptions, persistent myths and urban legends about likert scales and likert response formats and their antidotes. *Journal of Social Sciences*, 3(3):106–116, 2007.

[50] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[51] M.J. Chantler and J.P. Stoner. Automatic interpretation of sonar image sequences using temporal feature measures. *Oceanic Engineering, IEEE Journal of*, 22(1):47 –56, jan 1997.

[52] Tanya L. Chartrand and John A. Bargh. The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6):893–910, Jun 1999.

[53] Jixu Chen and Qiang Ji. A hierarchical framework for simultaneous facial activity tracking. In *FG*, pages 679–686, 2011.

[54] S.W. Chew, P. Lucey, S. Lucey, J. Saragih, J.F. Cohn, I. Matthews, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):1006 –1016, aug. 2012.

[55] S.W. Chew, S. Lucey, P. Lucey, S. Sridharan, and J.F. Conn. Improved facial expression recognition via uni-hyperplane classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2554–2561, 2012.

[56] Heeryon Cho, Toru Ishida, Naomi Yamashita, Rieko Inaba, Yumiko Mori, and Tomoko Koda. Culturally-situated pictogram retrieval. In *Proceedings of the 1st International Conference on Intercultural Collaboration*, pages 221–235, 2007.

[57] Ira Cohen, Ashutosh Garg, and Thomas S. Huang. Emotion recognition from facial expressions using multilevel HMM. *Neural Information Processing Systems (NIPS)*, 2000.

[58] Jeffrey Cohn and Karen Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1 – 12, March 2004.

[59] J.F. Cohn, L.I. Reed, Z. Ambadar, Jing Xiao, and T. Moriyama. Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 1, pages 610 – 616 vol.1, oct. 2004.

[60] Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, and Foldit players. Predicting protein structures with a multiplayer online game. *Nature*, 466:756–760, 2010.

[61] T. Cootes and C. Taylor. Active shape models - smart snakes. In *Proceedings of the British Machine Vision Conference*, pages 266–275, Leeds, UK, September 1992. Springer Verlag.

[62] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, pages 273–297, 1995.

[63] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs validated 3d human facial model. In *Proceedings of the SSPNET 2nd International Symposium on Facial Analysis and Animation*, FAA '10, pages 12–12, New York, NY, USA, 2010. ACM.

[64] Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[65] R. Cowie, E. Douglas-Cowie, B. Appolloni, J. Taylor, A. Romano, , and W Fellenz. What a neural net needs to know about emotion words. *Computational Intelligence and Applications*, pages 109–114, 1999. N. Mastorakis (Ed.), World Scientific & Engineering Society Press.

[66] R. Cowie, Douglas-Cowie E., and Cox C. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neural Networks*, 18:371–388, 2005.

[67] Roddy Cowie. Building the databases needed to understand rich, spontaneous human behaviour. In *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008.

[68] Roddy Cowie. *Philosophical Transactions of Royal Society B*, volume 364, chapter Perceiving emotion: towards a realistic understanding of the task, pages 3515–3525. Royal Society, 2009.

[69] Roddy Cowie, Gary McKeown, and Ceire Gibney. The challenges of dealing with distributed signs of emotion: theory and empirical evidence. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[70] David Cristinacce and Tim Cootes. Automatic feature localisation with constrained local models. *Pattern Recogn.*, 41(10):3054–3067, October 2008.

[71] Charles Darwin. *The Expression of the Emotions in Man and Animals*. Oxford University Press, 3rd edition, 2002.

[72] Dragoş Datcu and Léon Rothkrantz. Facial expression recognition in still pictures and videos using active appearance models: a comparison approach. In *Proceedings of the International Conference on Computer Systems and Technologies*, pages 1–6, New York, NY, USA, 2007. ACM.

[73] Alejandro H. De Mendoza. The problem of translation in cross-cultural research on emotion concepts (commentary on Choi & Han). *International Journal for Dialogical Science*, 3(1):241–248, 2008.

[74] A. Del Bue, X. Lladó, and L. Agapito. Non-rigid face modelling using shape priors. In S. Gong W. Zhao and X. Tang, editors, *Proceedings of the IEEE International Workshop on Analysis and Modelling of Faces and Gestures*, volume 3723 of *Lecture Notes in Computer Science*, pages 96–107. Springer-Verlag, 2005.

[75] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony : A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349– 365, July-September 2012.

[76] Laurence Devillers, Sarkis Abrilian, and Jean-Claude Martin. Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In *Affective Computing and Intelligent Interaction*, pages 519–526. Springer, 2005.

[77] Laurence Devillers and Jean-Claude Martin. Coding emotional events in audiovisual corpora. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[78] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883, 2011.

[79] C. Dietrich, F. Schwenker, and G. Palm. Classification of time series utilizing temporal and decision fusion. In Josef Kittler and Fabio Roli, editors, *Multiple Classifier Systems*, volume 2096 of *Lecture Notes in Computer Science*, pages 378–387. Springer Berlin / Heidelberg, 2001.

[80] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, October 1998.

[81] Gustav Lejeune Dirichlet. über die reduktion der positiven quadratischen formen mit drei unbestimmten ganzen zahle. *Journal für die Reine und Angewandte Mathematik*, 40:209227, 1850.

[82] S. D'Mello, A. Graesser, and R.W. Picard. Toward an affect-sensitive autotutor. *Intelligent Systems, IEEE*, 22(4):53 –61, july-aug. 2007.

[83] Gianluca Donato, Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 21(10):974–989, 1999.

[84] Fadi Dornaika and Franck Davoine. Simultaneous facial action tracking and expression recognition using a particle filter. In *Proceedings of the 10th IEEE International Conference on Computer Vision*, pages 1733–1738, Washington, DC, USA, 2005. IEEE Computer Society.

[85] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. Emotional speech: Towards a new generation of databases. *Speech Communication*, pages 33–60, 2003.

[86] Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in Neural Information Processing Systems*, pages 155–161, 1997. MIT Press.

[87] Micah Eckhardt and Rosalind Picard. A more effective way to label affective expressions. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[88] G. J. Edwards, C. J. Taylor, and T. F. Cootes. Interpreting face images using active appearance models. In *Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition*, page 300, Washington, DC, USA, 1998. IEEE Computer Society.

[89] Suzanne Eggins and Diana Slade. *Analysing Casual Conversation*. Equinox Publishing Ltd, London, 1997.

[90] P. Ekman. *Gesture, Speech, and Sign*, chapter Emotional and conversational nonverbal signals, pages 45–55. Oxford University Press, Oxford, 1979.

[91] P. Ekman. *Approaches to Emotion*, chapter Expression And The Nature Of Emotion, pages 319–344. Hillsdale New Jersey: Lawrence Erlbaum, 1984.

[92] P. Ekman. *Gesture, Speech and Sign*, chapter Emotional And Conversational Nonverbal Signals, pages 45–55. 1999.

[93] P. Ekman and W. V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:4898, 1969.

[94] P. Ekman and W. V. Friesen. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*. Prentice-Hall, Englewood Cliffs, N.J., 1975.

[95] P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978.

[96] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In J. Cole, editor, *Proceedings of the Nebraska Symposium on Motivation*, volume 19, pages 207–282. University of Nebraska Press, 1972.

[97] Paul Ekman. Basic emotions. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*. Wiley, Chichester, UK, 1999.

[98] Paul Ekman. *Philosophical Transactions of Royal Society B*, volume 364, chapter Darwin's contributions to our understanding of emotional expressions, pages 3449–3451. Royal Society, 2009.

[99] Rana el Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, volume 10, page 154, Washington, DC, USA, 2004. IEEE Computer Society.

[100] Rana el Kaliouby and Peter Robinson. *Real-time vision for HCI*, chapter Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures, pages 181–200. Springer, 2005.

[101] Rana el Kaliouby, Peter Robinson, and Simeon Keates. Temporal context and the recognition of emotion from facial expression. In *Proceedings of HCI International Conference*, June 2003.

[102] Rana Ayman el Kaliouby. *Mind-reading machines:automated inference of complex mental states*. PhD thesis, University of Cambridge, July 2005. UCAM-CL-TR-636.

[103] H. A. Elfenbein and N. Ambady. Is there an in-group advantage in emotion recognition? *Psychological Bulletin*, 128(2):243–249, 2002.

[104] H. A. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, 128(2):203–35, Mar 2002.

[105] R. A. Engle. Not channels but composite signals: Speech, gesture, diagrams and object demonstrations are integrated in multimodal explanations. In M.A. Gernsbacher and S.J. Derry, editors, *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, Mahwah, NJ, 1998. Erlbaum.

[106] Sergio Escalera, Rosa M. Martínez, Jordi Vitrià, Petia Radeva, and Teresa Anguera. Dominance detection in face-to-face conversations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 97–102, 2009.

[107] G. Fanelli, A.Yao, P.-L. Noel, J. Gall, and L. Van Gool. Hough forest-based facial expression recognition from video sequences. In *Proceedings of the International Workshop on Sign, Gesture and Activity (SGA)*, September 2010.

[108] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-D audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591–598, 2010.

[109] H. Fang and N. P. Costen. From rank-N to rank-1 face recognition based on motion similarity. In *Proceedings of the British Machine Vision Conference*, 2009.

[110] B. Fasel and J. Luettin. Recognition of asymmetric facial action unit activities and intensities. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 1, pages 1100–1103 vol.1, 2000.

[111] B. Fasel and J. Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003.

[112] B. Fehr and J. Russell. Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113:464, 1984.

[113] Robert Stephen Feldman, Pierre Philippot, and Robert J. Custrini. *Fundamentals of nonverbal behavior*, book; book/illustrated Social competence and nonverbal behavior, pages 329–350. Cambridge University Press, 1991. Includes bibliographical references and indexes.

[114] Xiaoyi Feng, Jie Cui, Matti Pietikäinen, and Abdenour Hadid. Real time facial expression recognition using local binary patterns and linear programming. In *Proceedings of the Mexican International Conference on Artificial Intelligence*, pages 328–336. Springer Berlin / Heidelberg, 2005.

[115] John Fiske. *Introduction to Communication Studies*. The John Fiske Collection. Routledge, 3 edition edition, November 14 2010.

[116] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. Ellsworth. The world of emotion is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

[117] M.G. Frank, P.N. Juslin, and J.A. Harrigan. *The New Handbook of Methods in Nonverbal Behaviour Research*, chapter Technical Issues in Recording Nonverbal Behaviour, pages 449–. Oxford University Press, 2005.

[118] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning*, pages 148–156, 1996.

[119] N. H. Frijda. *The Emotions*. Cambridge University Press, Cambridge, UK, 1986.

[120] Chris Frith. *Philosophical Transactions of Royal Society B*, volume 364, chapter Role of facial expressions in social interactions, pages 3453–3458. Royal Society, 2009.

[121] Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775 – 1787, 2009.

[122] K. F. Geisinger. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*, 6:304–312, 1994.

[123] Jeffrey M. Girard and Jeffrey F. Cohn. Criteria and metrics for thresholded AU detection. In *Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (BeFIT)*, Barcelona, Spain, November 2011.

[124] Linn Gjersing, John RM Caplehorn, and Thomas Clausen. Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. *BMC Medical Research Methodology*, 10(13), 2010.

[125] Roland Goecke. Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input. In *Proceedings of the NICTA-HCSNet Multimodal User Interaction Workshop*, pages 25–32, Darlinghurst, Australia, 2006. Australian Computer Society, Inc.

[126] Deborah Goren and Hugh R. Wilson. Quantifying facial expression recognition across viewing conditions. *Vision Research*, 46(8-9):1253–1262, Apr 2006.

[127] B. Goswami, Chi Ho Chan, J. Kittler, and B. Christmas. Local ordinal contrast pattern histograms for spatiotemporal, lip-based speaker authentication. In *Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on*, pages 1 –6, sept. 2010.

[128] Yves Grandvalet and Yoshua Bengio. Hypothesis testing for cross-validation. Technical Report 1285, Département d'informatique et recherche opérationnelle, Université de Montréal, 2006.

[129] Michael S. Gray, Javier R. Movellan, and Terrence J. Sejnowski. Dynamic features for visual speechreading: A systematic comparison. In *Advances in Neural Information Processing Systems*, volume 9, page 751, 1997.

[130] Samuel Green and Yanyun Yang. Commentary on coefficient alpha: A cautionary tale. *Psychometrika*, 74:121–135, 2009. 10.1007/s11336-008-9098-4.

[131] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 4, pages 1085–1088, 2007.

[132] M. Pantic H. Gunes, B. Schuller and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of the 1st International Workshop on Emotion Synthesis, Presentation, and Analysis in Continuous space, IEEE FG*, pages 827–834, Santa Barbara, California, USA, 21 March 2011. IEEE Press.

[133] Edward T. Hall. *The Silent Language*. Bantam Doubleday Dell Publishing Group, 1959.

[134] Judith A. Hall. *Encyclopedia of Social Psychology.*, chapter Nonverbal Cues and Communication, pages 626–28. SAGE, Thousand Oaks, CA, 2 Jul 2012.

[135] S. Haq and P. J. B. Jackson. Speaker-dependent audio-visual emotion recognition. In *Proc. Auditory-Visual Speech Processing.*, 2009.

[136] A. Hassan and R. I. Damper. Emotion recognition from speech using extended feature selection and a simple classifier. In *In Proceedings of 10th Annual of the International Speech Communication Association, Interspeech09*, page 24032406, Brighton, UK, 2009.

[137] Ali Hassan. *On automatic emotion classification using acoustic features*. PhD thesis, University of Southampton, June 2012.

[138] Andrew F. Hayes. *Statistical Methods For Communication Science*. Lawrence Erlbaum Associates, 2005.

[139] Lianghua He, Jianzhong Zhou, Die Hu, Cairong Zou, and Li Zhao. Boosted independent features for face expression recognition. In *Proceedings of the Second International Conference on Advances in Neural Networks*, pages 137–146, 2005.

[140] Dustin Hillard and Mari Ostendorf. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the Human Language Technology Conference / North American chapter of the Association for Computational Linguistics*, 2003.

[141] Ingrid Hopkins, Michael Gower, Trista Perez, Dana Smith, Franklin Amthor, F. Casey Wimsatt, and Fred Biasini. Avatar assistant: Improving social skills in students with an asd through a computer-based intervention. *Journal of Autism and Developmental Disorders*, 41:1543–1555, 2011. 10.1007/s10803-011-1179-z.

[142] Mohammed E. Hoque, Rana el Kaliouby, and Rosalind W. Picard. When human coders (and machines) disagree on the meaning of facial affect in spontaneous videos. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, Boston, Massachusetts, USA, April 2009.

[143] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. Models of attention in computing and communication: from principles to applications. *Commun. ACM*, 46(3):52–59, March 2003.

[144] Christopher K. Hsee, Elaine Hatfield, and Claude Chemtob. Assessments of the Emotional States of Others: Conscious Judgments Versus Emotional Contagion. *Journal of Social and Clinical Psychology*, 11(2):119–128, 1992.

[145] C. Humphrey. *Bilingual Women: Anthropological Approaches to Second Language Use*, chapter Casual Chat and Ethnic Identity: Women's Second-Language Use among Buryats in the USSR. Berg: Oxford, Mar 1993.

[146] Isabelle Hupont, Sandra Baldassarri, and Eva Cerezo. Facial emotional classification: from a discrete perspective to a continuous emotional space. *Pattern Analysis and Applications*, 16(1):41–54, 2013.

[147] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, unsupervised learning ica machine learning method 1999.

[148] Michael Isard and Andrew Blake. CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision (IJCV)*, 29(1):5–28, 1998.

[149] Rachael E. Jack, Caroline Blais, Christoph Scheepers, Philippe G. Schyns, and Roberto Caldara. Cultural confusions show that facial expressions are not universal. *Current Biology*, 19(18):1543 – 1548, 2009.

[150] Susan Jamieson. Likert scales: how to (ab)use them. *Medical Education*, 38(12):1217–1218, 2004.

[151] Fred E. Jandt. *An Introduction to Intercultural Communication*, chapter Nonverbal Communication, pages 120–145. Sage Publications, 4th edition edition, 2004.

[152] Laszlo A. Jeni, Daniel Takacs, and Andras Lorincz. High quality facial expression recognition in video streams using shape related information only. In *Proceedings of the First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (BeFIT)*, Barcelona, Spain, November 2011.

[153] Qiang Ji, Zhiwei Zhu, and P. Lan. Real-time nonintrusive monitoring and prediction of driver fatigue. *Vehicular Technology, IEEE Transactions on*, 53(4):1052 – 1068, july 2004.

[154] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11)*, pages 314–321, Santa Barbara, CA, USA, March 2011.

[155] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, pages 1–29, 2012.

[156] C.M. Jones and S.S. Dlay. The face as an interface: the new paradigm for HCI. *IEEE International Conference on Systems, Man, and Cybernetics*, 1:774–779 vol.1, 1999.

[157] Atul Kanaujia, Yuchi Huang, and Dimitris Metaxas. Emblem detections by tracking facial features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshop*, page 108, Washington, DC, USA, 2006. IEEE Computer Society.

[158] Ittipan Kanluan, Michael Grimm, and Kristian Kroschel. Audio-visual emotion recognition using an emotion space concept. In *Proceedings of the 16th European Signal Processing Conference*, Lausanne, Switzerland, August 2008.

[159] Kapka Kassabova. *Bulgaria*. New Holland Publishers, 2008.

[160] Abe Kazemzadeh, Sungbok Lee, and Shrikanth S. Narayanan. Fuzzy logic models for the meaning of emotion words. *IEEE Computational Intelligence Magazine*, 8(2):34–49, May 2013.

[161] A. Kendon. Some functions of gaze direction in social interaction. *Acta Psychologica*, 1967.

[162] Minyoung Kim and Vladimir Pavlovic. Structured output ordinal regression for dynamic facial emotion intensity prediction. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 649–662. Springer Berlin / Heidelberg, 2010.

[163] J Kittler. *Pattern Recognition and Signal Processing*, chapter Feature Set Search Algorithms, pages 41–60. Alphen aan den Rijn, The Netherlands: Sijthoff and Noordhoff, 1978.

[164] J. Kittler and F.M. Alkoot. Sum versus vote fusion in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 25(1):110 – 115, 2003.

[165] Mark L. Knapp. *Shared Experiences in Human Communication*, chapter Nonverbal Communication: Basic Perspectives, pages 91–106. Transaction Publishers, 1978.

[166] Mark L. Knapp and Judith A. Hall. *Nonverbal Communication in Human Interaction*. Cengage Learning, 2009.

[167] Ming Hsiao Ko, Geoff West, Svetha Venkatesh, and Mohan Kumar. Using dynamic time warping for online temporal fusion in multisensor systems. *Information Fusion*, 9(3):370 – 388, 2008. ¡ce:title¿Special Issue on Distributed Sensor Networks¡/ce:title¿.

[168] Tomoko Koda. Cross-cultural study of avatars' facial expressions and design considerations within Asian countries. In *Proceedings of the 1st International Conference on Intercultural Collaboration*, pages 207–220, 2007.

[169] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1940–1954, november 2010.

[170] Chen Y. & Chawla P. Krauss, R. M. *Advances in experimental social psychology*, chapter Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?, pages 389–450. Academic Press, San Diego, CA, 1996.

[171] J. B. Kruskal and M. Wish. Multidimensional scaling. Technical Report 07-011, Sage Publications, Beverly Hills and London, 1978. Sage University Paper series on Quantitative Application in the Social Sciences.

[172] Yuxuan Lan, Richard Harvey, Barry-John Theobald, Eng-Jon Ong, and Richard Bowden. Comparing visual features for lipreading. In *Proc. of International Conference on Auditory-visual Speech Processing*, Norwich, UK, Sept 2009.

[173] John T. Lanzetta and Robert E. Kleck. Encoding and decoding of nonverbal affect in humans. *Journal of Personality and Social Psychology*, 16(1):12–19, Sept 1970.

[174] Hyung-Soo Lee and Daijin Kim. Tensor-based aam with continuous variation estimation: Application to variation-robust face recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):1102 –1116, june 2009.

[175] V. Lee and G. Beattie. The rhetorical organization of verbal and nonverbal behavior in emotion talk. *Semiotica*, 120(1/2):39–92, 1998.

[176] Jaakko Lehtonen. Non-verbal aspects of impromptu speech. In Nils Erik Enkvist, editor, *Problems in the Linguistic Study of Impromptu Speech*, Abo, Finland, November 20-22 1981. ERIC Clearinghouse.

[177] Wenhui Liao, Weihong Zhang, Zhiwei Zhu, Qiang Ji, and Wayne D. Gray. Toward a decision-theoretic framework for affect recognition and user assistance. *International Journal of Human-Computer Studies*, 64(9):847 – 873, 2006.

[178] J. Lien, T. Kanade, J. F. Cohn, C. Li, and A. Zlochower. Subtly different facial expression recognition and expression intensity estimation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, page 853, Washington, DC, USA, 1998. IEEE Computer Society.

[179] J. J. Lien, T. Kanade, J. F. Cohn, and Ching-Chung Li. Automated facial expression recognition based on FACS action units. In *Proceedings of the 3rd IEEE International Conference. Automatic Face and Gesture Recognition*, pages 390–395, Apr 1998.

[180] Rensis Likert. A technique for the measurement of attitudes. *Archives of Psychology*, 140:155, 1932.

[181] Da Yu Lin, Shao-Zhong Zhang, Eric Block, and Lawrence C. Katz. Encoding social signals in the mouse main olfactory bulb. *Nature*, 434:470–477, 2005.

[182] Jackson Liscombe, Jennifer Venditti, and Julia Hirschberg. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Eurospeech*, 2003.

[183] Zicheng Liu and Zhengyou Zhang. Robust head motion computation by taking advantage of physical properties. In *Proceedings of the Workshop on Human Motion*, pages 73–77, 2000.

[184] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Efficient online subspace learning with an indefinite kernel for visual tracking and recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1624–1636, October 2012.

[185] S. Liwicki, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Euler principal component analysis. *International Journal of Computer Vision*, 2012. (in press).

[186] MichaelF. Lorber, SusanG. O'Leary, and KimberlyT. Kendziora. Mothers' overreactive discipline and their encoding and appraisals of toddler behavior. *Journal of Abnormal Child Psychology*, 31(5):485–494, 2003.

[187] Xiaoguang Lu. *Three-dimensional face recognition across pose and expression.* PhD thesis, Michigan State University, East Lansing, MI, USA, 2006. AAI3236364.

[188] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of Imaging understanding workshop*, 1981.

[189] Patrick Lucey, Jeffrey Cohn, Simon Lucey, and Iain Matthews. Automatically detecting pain using facial actions. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[190] J. Luettin, N.A. Thacker, and S.W. Beet. Speechreading using shape and intensity information. *Fourth International Conference on Spoken Language (ICSLP 96)*, 1:58–61 vol.1, Oct 1996.

[191] Donald M MacKay. *Non-verbal Communication*, chapter Formal analysis of communicative processes, pages 3–. Cambridge U. Press, 1972.

[192] Anmol P. Madan. *Thin Slices of Interest.* PhD thesis, MIT, 2003.

[193] Anmol P. Madan, Ron Caneel, and Alex Pentland. Voices of attraction. In *Proceedings of Augmented Cognition, HCI*, Las Vegas, NV, 2005.

[194] Haim Mano and Richard L Oliver. Assessing the dimensionality and structure of the consumption experience: Evaluation, feeling, and satisfaction. *Journal of Consumer Research*, 20(3):451–66, 1993.

[195] R. A. Mar and K. Oatley. The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science*, 13:173–192, 2008.

[196] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, pages 570–576, 1998. MIT Press.

[197] Abigail A. Marsh, Hillary Anger Elfenbein, and Nalini Ambady. Nonverbal "accents": Cultural differences in facial expressions of emotion. *Psychological Science*, 14(4):373–376, 2003.

[198] DC Martin. *Clinical Methods: The History, Physical, and Laboratory Examinations*, chapter The Mental Status Examination. Number 207. Butterworths, Boston, 3rd edition edition, 1990.

[199] D.W. Massaro and J.W Ellison. Perceptual recognition of facial affect: Cross-cultural comparisons. *Memory & Cognition*, 24(6):812–822, 1996.

[200] T. Masuda, P. C. Ellsworth, B. Mesquita, J. Leu, S. Tanida, and E. van de Veerdonk. Placing the face in context: Cultural differences in the perception of facial emotion. *Journal of Personality and Social Psychology*, 94:365–381, 2008.

[201] David Matsumoto. *The SAGE Handbook of Nonverbal Communication*, chapter Culture and Nonverbal Behavior, pages 219–236. Sage Publications, Inc, 2006.

[202] David Matsumoto, Paul Ekman, and Alan Fridlund. *Practical Guild to Using Video in the Behavioral Sciences*, chapter Analyzing Nonverbal Behavior, pages 153–165. John Wiley & Sons, 1991.

[203] David Matsumoto, Jeff Leroux, Carinda Wilson-Cohn, Jake Raroque, Kristie Kooken, Paul Ekman, Nathan Yrizarry, Sherry Loewinger, Hideko Uchida, Albert Yee, Lisa Amo, and Angeline Goh. A new test to measure emotion recognition ability: Matsumoto and Ekmans Japanese and Caucasian brief affect recognition test (JACBART). *Journal of Nonverbal Behavior*, 24(3):179209, 2000.

[204] David Matsumoto, Seung H. Yoo, and Johnny Fontaine. Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of Cross-Cultural Psychology*, 39(1):55–74, January 2008.

[205] Iain Matthews, Tim Cootes, Stephen Cox, Richard Harvey, and J. Andrew Bangham. Lipreading using shape, shading and scale. In *Proceedings of International Conference on Auditory-Visual Speech Processing*, pages 73–78, Sydney, Australia, December 1998.

[206] Anjanie McCarthy, Kang Lee, Shoji Itakura, and Darwin W Muir. Cultural display rules drive eye gaze during thinking. *Journal of Cross-Cultural Psychology*, 37(6):717–722, 2006.

[207] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang. Automatic analysis of multimodal group actions in meetings. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):305 –317, march 2005.

[208] Margaret McRorie and Ian Sneddon. Real emotion is dynamic and interactive. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, volume 4738/2007, pages 759–760, 2007.

[209] A. Meng, P. Ahrendt, J. Larsen, and L.K. Hansen. Temporal feature integration for music genre classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5):1654 –1664, july 2007.

[210] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior research methods*, 37(4):626–630, 2005.

[211] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti. Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. *Psychophysiology*, 2010.

[212] Stephen Moore and Richard Bowden. Automatic facial expression recognition using boosted discriminatory classifiers. In *Proceedings of the 3rd IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pages 71–83, 2007.

[213] Stephen Moore and Richard Bowden. The effects of pose on facial expression recognition. In *Proceedings of the British Machine Vision Conference*, pages 1–11, London, 2009.

[214] Leo S. Morales. *Assessing Patient Experiences with Assessing Healthcare in Multi-Cultural Settings*. PhD thesis, RAND Graduate School, 2001.

[215] Louis-Philippe Morency. Computational study of human communication dynamic. In *Proceedings of the 2011 joint ACM workshop on Human Gesture and Behavior Understanding*, pages 13–18, New York, NY, USA, 2011. ACM.

[216] Edward R. Morrison, Lisa Gralewski, Neill Campbell, and Ian S. Penton-Voaka. Facial movement varies by sex and is related to attractiveness. *Evolution and Human Behavior*, 28(3):186–192, 2007.

[217] Michael T. Motley. Consciousness and intentionality in communication: A preliminary model and methodological approaches. *Western Journal of Speech Communication*, 50(1):3, 1986.

[218] Emily Mower, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Interpreting ambiguous emotional expressions. In *Proceedings of the International Conference on Affective Computing & Intelligent Interaction (ACII)*, Amsterdam, Sept 2009.

[219] Rhyohei Nakatsu. Nonverbal information recognition and its application to communications. In *Proceedings of the 6th ACM International Conference on Multimedia*, pages 2–9, New York, NY, USA, 1998. ACM.

[220] Sano Natsuki, Suzuki Hideo, and Koda Masato. A robust ensemble learning using zero-one loss function. *Journal of the Operations Research Society of Japan*, 51(1):95–110, 2008-03.

[221] Ara V. Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Applied Signal Processing*, 2002(1):1274–1288, 2002.

[222] Janak Singh Negi. The role of teachers non-verbal communication in elt classroom. *Journal of NELTA*, 14(1&2):101–110, December 2009.

[223] M. A. Nicolaou, H. Gunes, and M. Pantic. Output-associative RVM regression for dimensional and continuous emotion prediction. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, pages 16–23, Santa Barbara, CA, USA, March 2011.

[224] M.A. Nicolaou, H. Gunes, and M. Pantic. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *Affective Computing, IEEE Transactions on*, 2(2):92–105, 2011.

[225] Tore A. Nielsen, Daniel Deslauriers, and George W. Baylor. Emotions in dream and waking event reports. *Dreaming: Journal of the Association for the Study of Dreams*, 1(4):287–300, Dec 1991.

[226] Geoff Norman. Likert scales, levels of measurement and the laws of statistics. *Advances in Health Sciences Education*, 15:625–632, 2010. 10.1007/s10459-010-9222-y.

[227] C. Oertel, F. Cummins, N. Campbell, J. Edlund, and P. Wagner. D64: A corpus of richly recorded conversational interaction. In *Proceedings of the Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (LREC)*, Valetta, Malta, 2010.

[228] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatiotemporal localization and categorization of human actions in unsegmented image sequences. *IEEE Transactions on Image Processing*, 20(4):1126–1140, April 2011.

[229] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(7):971–987, 2002.

[230] E-J. Okwechime, D. Ong, A. Gilbert, and R. Bowden. Visualisation and prediction of conversation interest through mined social signals. In *IEEE International Workshop on Social Behavior Analysis*, 2011.

[231] E. Ong, Y. Lan, B. Thobald, R. Harvey, and R. Bowden. Robust facial feature tracking using multiscale biased linear predictors. In *Proceedings of the International Conference on Computer Vision*, 2009.

[232] Eng-Jon Ong and R. Bowden. Learning temporal signatures for lip reading. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 958 –965, nov. 2011.

[233] Charles Ochieng Owuor. *Implications of using Likert data in multiple regression analysis*. PhD thesis, University of British Columbia, 2001.

[234] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(2):433–449, April 2006.

[235] M Pantic and A Vinciarelli. Implicit human centered tagging. *IEEE Signal Processing Magazine*, 26(6):173–180, 2009.

[236] Maja Pantic. *Philosophical Transactions of Royal Society B*, volume 364, chapter Machine analysis of facial behaviour: Naturalistic and dynamic behaviour, pages 3505–3513. Royal Society, 2009.

[237] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S. Huanag. Human-centred intelligent human computer interaction (HCI2): how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168–187, 2008.

[238] Ioannis Patras and Maja Pantic. Particle filtering with factorized likelihoods for tracking facial features. In *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 97–104, Seoul, Korea, May 17-19 2004. IEEE Computer Society.

[239] W. B. Pearce and K Kang. *Cross-cultural Adaptation*, volume IX, chapter Conceptual migrations: Understanding travelers tales for cross-cultural understanding, pages 20–40. Sage, Beverly Hills, CA, 1987.

[240] A. Pentland. Social signal processing [exploratory dsp]. *Signal Processing Magazine, IEEE*, 24(4):108 –111, july 2007.

[241] Susan B. Perlman, James P. Morris, Brent C. Vander Wyk, Steven R. Green, Jaime L. Doyle, and Kevin A. Pelphrey. Individual differences in personality predict how people look at faces. *PLoS ONE*, 4(6):e5952, 06 2009.

[242] J. Perner. *Developmental psychology: Achievements & prospects*, chapter Theory of Mind, pages 205–230. Psychology Press, Hove, East Sussex, 1999.

[243] S. Petridis, A. Asghar, and M. Pantic. Classifying laughter and speech using audio-visual feature prediction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, USA, March 2010.

[244] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic. Static vs. dynamic modelling of human nonverbal behaviour from multiple cues and modalities. In *Proceedings of the ACM Interational Conference on Multimodal Interfaces*, Cambridge, USA, November 2009.

[245] S. Petridis and M. Pantic. Audiovisual discrimination between speech and laughter: Why and when visual information might help. *IEEE Transactions on Multimedia*, 13(2):216–234, April 2011.

[246] Stavros Petridis and Maja Pantic. Audiovisual laughter detection based on temporal features. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 37–44, New York, NY, USA, 2008. ACM.

[247] Julien Peyras, Adrien Bartoli, and Samir Khoualed. Pools of AAMs: Towards automatically fitting any face image. In *Proceedings of the Nineth British Machine Vision Conference*, Leeds, UK, September 2008.

[248] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In *Proceedings of the Workshop on Socially Intelligent Surveillance and Monitoring*, 2011.

[249] Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikäinen. Recognising spontaneous facial micro-expressions. In *Proceedings of the International Conference on Computer Vision*, 2011.

[250] I. Poggi and F D'Errico. Social signals: A psychological perspective. In *In Proceedings of Computer Analysis of Human Behavior*, pages 185–225, 2011.

[251] Norman Poh and Josef Kittler. *Multimodal Signal Processing: Theory and Applications for Human-Computer Interaction*, chapter Multimodal Information Fusion, pages 153–169. Academic Press, London, 2010.

[252] G. Potamianos, J. Luettin, and C. Neti. Hierarchical discriminant features for audio-visual lvcsr. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 165–168, Salt Lake City, Utah, USA, 2001.

[253] Fernando Poyatos, editor. *Nonverbal Communication and Translation.* Amsterdam/Philadelphia, John Benjamins, 1997.

[254] J. Raddick, G. L. Bracey, and P. L. Gay. Motivations of citizen scientists participating in galaxy zoo: A more detailed look. In *American Astronomical Society Meeting Abstracts #215*, volume 42 of *Bulletin of the American Astronomical Society*, page 509, January 2010.

[255] F. Ramseyer and W. Tschacher. *Simultaneity: Temporal structures and observer perspectives.*, chapter Synchrony in dyadic psychotherapy sessions, pages 329–347. World Scientific, Singapore, 2008.

[256] Jeffrey L. Rasmussen. Analysis of likert-scale data: A reinterpretation of gregoire and driver. *Psychological Bulletin*, 105(1):167–170, Jan 1989.

[257] Dennis Reidsma. *Annotations and subjective machines of annotators, embodied agents, users, and other humans.* PhD thesis, University of Twente, Enschede, October 2008.

[258] Dennis Reidsma, Dirk Heylen, and H. J. A. op den Akker. On the contextual analysis of agreement scores. In Jean-Claude Martin, Patrizia Paggio, Michael Kipp, and Dirk Heylen, editors, *Proceedings of the Workshop on Multimodal Corpora (LREC)*, pages 52–55. ELRA, May 2008.

[259] Dennis Reidsma, Anton Nijholt, Wolfgang Tschacher, and Fabian Ramseyer. Measuring multimodal synchrony for human-computer interaction. In A. Sourin, editor, *Proceedings of the International Conference on CYBERWORLDS 2010*, pages 67–71, Los Alamitos, October 2010. IEEE Computer Society Press. Synchrony, nonverbal communication, measurement, virtual humans, HCI.

[260] Dennis Reidsma and Rieks op den Akker. Exploiting subjective annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistic*, Manchester, UK, August 2008.

[261] William Revelle and Richard Zinbarg. Coefficients alpha, beta, omega, and the glb: Comments on sijtsma. *Psychometrika*, 74:145–154, 2009. 10.1007/s11336-008-9102-z.

[262] Daniel C. Richardson, Rick Dale, and Natasha Z. Kirkham. The art of conversation is coordination: Common ground and the coupling of eye movements during dialogue. *Psychological Science*, 18:407–413(7), May 2007.

[263] N. Rose. A comparison of single and multi-class classifiers for facial expression classification. In *Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, page 175, Nov. 28-Dec. 1 2006.

[264] M. Rosenblum, Y. Yacoob, and L.S. Davis. Human expression recognition from motion using a radial basis function network architecture. *IEEE Transactions on Neural Networks*, 7(5):1121–1138, September 1996.

[265] Richard M. Rozelle, Daniel Druckman, and James C. Baxter. *The Handbook of Communication Skills*, chapter Non-verbal behavior as communication, pages 67–102. Routledge, 1st July 2006.

[266] O. Rudovic and M. Pantic. Shape-constrained gaussian process regression for facial-point-based head-pose normalization. In *Proceedings of IEEE Intl Conf. on Computer Vision (ICCV 2011)*, pages 1495–1502, November 2011.

[267] O. Rudovic, I. Patras, and M. Pantic. Regression-based multi-view facial expression recognition. In *Proceedings of Int'l Conf. Pattern Recognition (ICPR'10)*, pages 4121–4124, Istanbul, Turkey, August 2010.

[268] O. Rudovic, V. Pavlovic, and M. Pantic. Kernel conditional ordinal random fields for temporal segmentation of facial action units. In *Proceedings of the 12th European Conference on Computer Vision (ECCV-W'12). Florence, Italy*, October 2012.

[269] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Multi-output Laplacian Dynamic Ordinal Regression for Facial Expression Recognition and Intensity Estimation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012.

[270] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, September 2007.

[271] T. Saitoh and R. Konishi. Lip reading based on sampled active contour model. In *Proceedings of the International Conference on Image Analysis and Recognition*, pages 507–515, 2005.

[272] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, 18:401–409, 1969.

[273] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), Special Session: 3D Facial Behavior Analysis and Understanding*, pages 406–413, Santa Barbara, CA, USA, March 2011.

[274] Disa A. Sauter, Frank Eisner, Paul Ekman, and Sophie K. Scott. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 2009.

[275] Abraham Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[276] Arman Savran, Bulent Sankur, and M. Taha Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774 – 784, 2012. ¡ce:title¿3D Facial Behaviour Analysis and Understanding¡/ce:title¿.

[277] K. R. Scherer. *Handbook of cognition and emotion*, chapter Appraisal theory, pages 637–663. Wiley, Chichester, 1999.

[278] B. Schölkopf, A. Smola, R. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

[279] M. Schroder, S. Pammi, H. Gunes, M. Pantic, M. Valstar, R. Cowie, G. McKeown, D. Heylen, M. ter Maat, F. Eyben, B. Schuller, M. Wollmer, E. Bevacqua, C. Pelachaud, and E. de Sevin. Come and have an emotional workout with sensitive artificial listeners! In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Santa Barbara, California, USA, 2011.

[280] Marc Schröder, R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, and M. Sawey. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*, pages 19–24, Belfast, 2000. Textflow.

[281] Bjrn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(910):1062 – 1087, 2011. ¡ce:title¿Sensing Emotion and Affect - Facing Realism in Speech Processing¡/ce:title¿.

[282] R. Seguier, N. Cladel, C. Foucher, and D. Mercier. Lipreading with spiking neurons: One pass learning. In *Proceedings of the 10th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, page 397, Plzen, 2002.

[283] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):993 –1005, aug. 2012.

[284] Dino Seppi, Anton Batliner, Bjrn Schuller, Stefan Steidl, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Vered Aharonson, and Fondazione Bruno Kessler. Patterns, prototypes, performance: Classifying emotional user states. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, Brisbane, Australia, 2008.

[285] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[286] P. Shaver, J. Schwartz, D. Kirson, and C. O'connor. Emotion knowledge: further exploration of a prototype approach. *Journal of personality and social psychology*, 52(6):1061, 1987.

[287] Tim Sheerman-Chase, Eng-Jon Ong, and Richard Bowden. Feature selection of facial displays for detection of non verbal communication in natural conversation. In *Proceedings of the IEEE International Workshop on Human-Computer Interaction*, Kyoto, Oct 2009.

[288] Tim Sheerman-Chase, Eng-Jon Ong, and Richard Bowden. Non-linear predictors for facial feature tracking across pose and expression. In *IEEE Conference on Automatic Face and Gesture Recognition, Shanghai*, 2013. (In press).

[289] Klaas Sijtsma. On the use, the misuse, and the very limited usefulness of cronbachs alpha. *Psychometrika*, 74(1):107–120, March 2009.

[290] Klaas Sijtsma. Reliability beyond theory and into practice. *Psychometrika*, 74(1):169–173, March 2009.

[291] Paul Smith, Niels da Vitoria Lobo, and Mubarak Shah. Temporal boost for event recognition. In *Proceedings of the 10th IEEE International Conference on Computer Vision*. IEEE, October 2005.

[292] Tal Sobol-Shikler and Peter Robinson. Classification of complex information: inference of co-occurring affective states from their expressions in speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(7):1284–1297, July 2010.

[293] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *Performance Evaluation*, 4304(c):10151021, 2006.

[294] T. Starner and A. Pentland. Real-time American sign language recognition from video using hidden Markov models. In *Proceedings of the International Symposium on Computer Vision*, page 5B Systems and Applications, 1995.

[295] Michael Stubbs. *Discourse Analysis: the sociolinguistic analysis of natural language*. Basil Blackwell, Oxford, 1983.

[296] Xiaofan Sun, Jeroen Lichtenauer, Michel Valstar, Anton Nijholt, and Maja Pantic. A multimodal database for mimicry analysis. In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction, Part I*, volume 6974 of *Lecture Notes in Computer Science*, pages 367–376, Berlin, Germany, October 2011. Springer Verlag.

[297] Xiaofan Sun, Anton Nijholt, and Maja Pantic. Towards mimicry recognition during human interactions: Automatic feature selection and representation. In Antonio Camurri and Cristina Costa, editors, *Intelligent Technologies for Interactive Entertainment*, volume 78 of *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, pages 160–169, Heidelberg, Germany, September 2012. Springer Verlag.

[298] Yafei Sun, Nicu Sebe, Michael Lew, and Theo Gevers. Authentic emotion detection in real-time video. In *Proceedings of the International Workshop on Human Computer Interaction*, volume 3058, pages 94–104, Prague, 2004.

[299] Jaewon Sung, Takeo Kanade, and Daijin Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *Int. J. Comput. Vision*, 80(2):260–274, November 2008.

[300] Sima Taheri, Pavan K. Turaga, and Rama Chellappa. Towards view-invariant expression analysis using analytic shape manifolds. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 306–313, 2011.

[301] Fangqi Tang and Benzai Deng. Facial expression recognition using AAM and local facial features. In *Proceedings of the Third International Conference on Natural Computation*, volume 2, pages 632–635, August 2007.

[302] A. Tarasov, C. Cullen, and S Delany. Using crowdsourcing for labeling emotional speech assets. In *Proceedings of the W3C Workshop on Emotion ML*, Paris, France, 2010.

[303] Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S. Huang, Xutao Lv, and Tony X. Han. Emotion recognition from an ensemble of features. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 872–877, 2011.

[304] David Tax, Emile Hendriks, Michel F. Valstar, and Maja Pantic. The detection of concept frames using clustering multi-instance learning. In *Proceedings of International Conference on Pattern Recognition*, pages 2917–2921, 2010.

[305] A Terracciano, M Merritt, AB Zonderman, and MK Evans. Personality traits and sex differences in emotions recognition among African Americans and Caucasians. *Annals of New York Academy of Sciences*, 1000:309–312, 2003.

[306] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical report, Carnegie Mellon University Technical Report CMU-CS-91-132, April 1991.

[307] Horst Treiblmaier and Peter Filzmoser. Benefits from using continuous rating scales in online survey research. In *International Conference on Information Systems (ICIS)*, 2011.

[308] K.P. Truong. *How does real affect affect affect recognition in speech?* PhD thesis, University of Twente, Enschede, September 2009.

[309] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. T. Pavlidis, M. G. Frank, and P. Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision (IJCV)*, 71(2):197–214, 2007.

[310] G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. Robust and efficient parametric face alignment. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1847–1854, November 2011.

[311] H.D. Vakayallapati, K.R Anne, and K. Kyamakya. Emotion recognition from decision level fusion of visual and acoustic features using hausdorff classifier. In *Fifth International Conference on Information Processing (ICIP-2011)*, 2011.

[312] M F Valstar and M Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, pages 65–70, Malta, May 2010.

[313] M.F. Valstar, M. Mehu, Bihan Jiang, M. Pantic, and K. Scherer. Meta-analysis of the first facial expression recognition challenge. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):966 –979, aug. 2012.

[314] Michel F. Valstar, Bihan Jiang, Marc Méhu, Maja Pantic, and Klaus Scherer. The first facial expression recognition and analysis challenge. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2011.

[315] Michel F. Valstar, Marc Méhu, Marcello Mortillaro, Maja Pantic, and Klaus Scherer. Meta-analysis of challenge slides. In *First Facial Expression Recognition and Analysis Challenge (FERA2011)*, 2011.

[316] Michel F. Valstar and Maja Pantic. Fully automatic recognition of the temporal phases of facial actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(1):28–43, February 2012.

[317] Michel F. Valstar, Maja Pantic, Zara Ambadar, and Jeffrey F. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *Proceedings of the 8th International Conference on Multimodal interfaces*, pages 162–170, New York, NY, USA, 2006. ACM.

[318] V Vapnik. *Statistical Learning Theory*. John Wiley, New York, NY, 1998.

[319] P.K. Varshney. Multisensor data fusion. *Electronics Communication Engineering Journal*, 9(6):245 –253, dec 1997.

[320] Suzane Vassallo, Sian L. Cooper, and Jacinta M. Douglas. Visual scanning in the recognition of facial affect: Is there an observer sex difference? *Journal of Vision*, 9(3):1–10, 3 2009.

[321] Rudolph F. Verderber and Kathleen S. Verderber. *Communicate!*, chapter Communicating Through Non-verbal Behaviours, page 77. Cengage Learning Editores, 2007.

[322] Philippe Verduyn, Ellen Delvaux, Hermina Van Coillie, Francis Tuerlinckx, and Iven Van Mechelen. Predicting the duration of emotional experience: Two experience sampling studies. *Emotion*, 9(1):83–91, 2009.

[323] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction*, 2009.

[324] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, December 2008.

[325] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2002.

[326] Michiel Visser, Mannes Poel, and Anton Nijholt. Classifying visemes for automatic lipreading. In *Proceedings of the Second International Workshop on Text, Speech and Dialogue*, pages 349–352, London, UK, 1999. Springer-Verlag.

[327] Bin Wang and Wenkai Lu. An in-depth comparasion on FastICA, CuBICA and IC-FastICA. In *Proceedings of the Advances in Natural Computation*, Lecture Notes in Computer Science, pages 410–414, 2005.

[328] Jun Wang, Lijun Yin, Xiaozhou Wei, and Yi Sun. 3D facial expression recognition based on primitive surface feature distribution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1399–1406, 2006.

[329] Eric W. Weisstein. Correlation coefficient. From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/CorrelationCoefficient.html.

[330] B. L. Welch. The generalization of "students problem when several different population variances are involved. *Biometrika*, 34:2835, 1947.

[331] Sheida White. Backchannels across cultures: A study of Americans and Japanese. *Language in Society*, 18(1):59–76, 1989.

[332] M. Wiener, S. Devoe, S. Rubinow, and J. Geller. Nonverbal behavior and nonverbal communication. *Psychological Review*, 79(4):185–21, 1972.

[333] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie, and Roddy Cowie. Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, pages 597–600, 2008.

[334] Martin Wöllmer, Florian Eyben, Björn Schuller, Ellen Douglas-Cowie, and Roddy Cowie. Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, ISSN 1990-9772, pages 1595–1598, Brighton, UK, 2009.

[335] Tingfan Wu, N.J. Butko, P. Ruvolo, J. Whitehill, M.S. Bartlett, and J.R. Movellan. Multilayer architectures for facial action unit recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(4):1027 –1038, aug. 2012.

[336] Xu Xiang-min, Mao Yun-feng, Xiong Jia-ni, and Zhou Feng-le. Classification performance comparison between rvm and svm. In *Anti-counterfeiting, Security, Identification, 2007 IEEE International Workshop on*, pages 208 –211, april 2007.

[337] Jimei Yang, Shengcai Liao, and Stan Z. Li. Automatic partial face alignment in nir video sequences. In Massimo Tistarelli and MarkS. Nixon, editors, *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 249–258. Springer Berlin Heidelberg, 2009.

[338] Peng Yang, Qingshan Liu, and D.N. Metaxas. Rankboost with l1 regularization for facial expression recognition and intensity estimation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1018 –1025, 29 2009-oct. 2 2009.

[339] Songfan Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *Proceedings of the International Conference on Automatic Face & Gesture Recognition and Workshops*, pages 866 – 871, Santa Barbara, CA, 2011.

[340] Yong Yang, Guoyin Wang, and Hao Kong. Self-learning facial emotional feature selection based on rough set theory. *Mathematical Problems in Engineering*, 2009. Article ID 802932.

[341] V Yngve. On getting a word in edgeways. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, Chicago, 1970. Chicago Linguistic Society.

[342] A Zara, V Maffiolo, J Martin, and L Devillers. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. *Affective Computing and Intelligent Interaction*, pages 464–475, 2007.

[343] Z. Zeng, Y. Fu, G.I. Roisman, Z. Wen, Y. Hu, and T.S. Huang. Spontaneous emotional facial expression detection. *Journal of Multimedia*, 1(5):1–8, 2006.

[344] Zhihong Zeng, M. Pantic, G.I. Roisman, and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, jan. 2009.

[345] Jianke Zhu, Luc Van Gool, and S.C.H. Hoi. Unsupervised face alignment by robust nonrigid mapping. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1265 –1272, 29 2009-oct. 2 2009.

[346] Xiangxin Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2879 –2886, june 2012.

[347] Wobbe P. Zijlstra, L. Andries van der Ark, and Klaas Sijtsma. Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics*, 36(2):186–212, 2011.